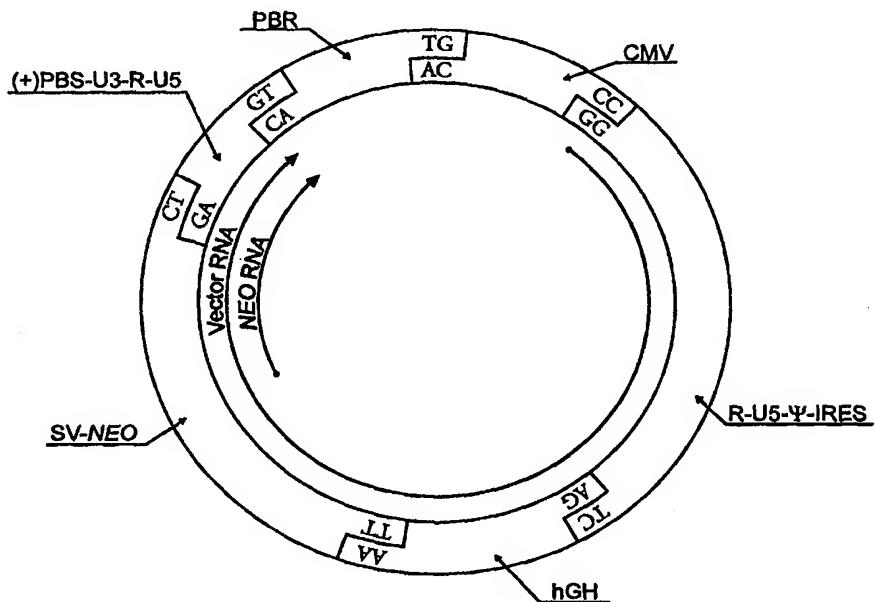




## INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification <sup>6</sup> :  C12N 15/86, 15/10		A1	(11) International Publication Number: <b>WO 98/38326</b>
			(43) International Publication Date: 3 September 1998 (03.09.98)
(21) International Application Number: PCT/US98/03918  (22) International Filing Date: 28 February 1998 (28.02.98)		(81) Designated States: AU, CA, JP, US, European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).	
(30) Priority Data: 60/070,910 28 February 1997 (28.02.97) US		Published <i>With international search report. Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.</i>	
(71) Applicant (for all designated States except US): NATURE TECHNOLOGY CORPORATION [US/US]; 109 South 54th Street, Omaha, NE 68132 (US).			
(72) Inventors; and (75) Inventors/Applicants (for US only): HODGSON, Clague, P. [US/US]; 109 South 54th Street, Omaha, NE 68132 (US). ZINK, Mary, Ann [US/US]; 109 South 54th Street, Omaha, NE 68132 (US). XU, Guoping [CN/US]; 109 South 54th Street, Omaha, NE 68132 (US).			
(74) Agent: MCCORMACK, Myra, H.; Mueting, Raasch & Gebhardt, P.A., P.O. Box 581415, Minneapolis, MN 55458-1415 (US).			

(54) Title: SELF-ASSEMBLING GENES, VECTORS AND USES THEREOF



## (57) Abstract

The invention relates to a method for directing the self-assembly of a gene or gene assembly having three and preferably six or more fragments in a directionally and spatially ordered fashion to produce a gene, gene vector or large nucleic acid molecule. The method can be used to create libraries, such as combinatorial libraries. In another embodiment of the invention a vector is described for the incorporation and screening of endogenous mouse promoter elements for the identification of cell-specific promoters.

***FOR THE PURPOSES OF INFORMATION ONLY***

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece			TR	Turkey
BG	Bulgaria	HU	Hungary	ML	Mali	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MN	Mongolia	UA	Ukraine
BR	Brazil	IL	Israel	MR	Mauritania	UG	Uganda
BY	Belarus	IS	Iceland	MW	Malawi	US	United States of America
CA	Canada	IT	Italy	MX	Mexico	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NE	Niger	VN	Viet Nam
CG	Congo	KE	Kenya	NL	Netherlands	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NO	Norway	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	NZ	New Zealand		
CM	Cameroon			PL	Poland		
CN	China	KR	Republic of Korea	PT	Portugal		
CU	Cuba	KZ	Kazakhstan	RO	Romania		
CZ	Czech Republic	LC	Saint Lucia	RU	Russian Federation		
DE	Germany	LI	Liechtenstein	SD	Sudan		
DK	Denmark	LK	Sri Lanka	SE	Sweden		
EE	Estonia	LR	Liberia	SG	Singapore		

**SELF-ASSEMBLING GENES, VECTORS AND USES THEREOF****Field of the Invention**

5 This invention relates to the construction and usage of synthetic genes for genetic engineering and gene therapy.

**Background of the invention**

This application claims the benefit of a provisional application U.S. Serial No. 10 60/070,910, filed on February 28, 1997, entitled "Self-Assembling Genes."

Recombination at the genetic level is important for generating diversity and adaptive change within genomes of virtually all organisms. Recombinant DNA technology is based upon simple 'cut-and-paste' methods for manipulating nucleic acid molecules *in vitro*. The pieces of genetic material, or DNA are first digested with a restriction endonuclease 15 enzyme which recognizes specific sequences within the DNA. After preparation of two or more pieces of DNA, the ends of the DNA are further manipulated, if necessary, to make them compatible for ligation or joining together. DNA ligase, together with adenosine triphosphate (ATP) is added to the genes, ligating them back together. The genetic assembly containing an origin of DNA replication and a selectable gene is then inserted into a living 20 cell, is grown up, and is positively selected to yield a pure culture capable of providing high yields of individual recombinant DNA molecules, or their products such as RNA or protein.

Significant improvements have been made to this technology over the last two and a half decades. Numerous enzymes, end-linkers and adapter molecules have been made commercially available, which facilitate in the construction of recombinant DNA molecules. 25 By using two restriction enzymes with different single-stranded termini or blunt ends, it is possible to directionally assemble genes (forced cloning). This reduces the amount of screening required to determine orientation. Procedures have been automated for synthesis of single-stranded gene fragments up to 200 or more nucleotides in length by means of phosphoramidite chemistry, and the instrumentation is readily available through Applied Biosystems, Inc., Foster City, CA. Such single-stranded fragments can be joined by annealing overlapping complimentary phosphorylated strands, and by enzymatically filling in the ends with DNA polymerase and DNA precursors. In this way, multiple, overlapping, 30 single-stranded fragments can be assembled into a larger, double-stranded superstructure.

Whole genes have been synthesized by similar methods. However, it becomes increasingly difficult to use synthetic DNA strands when making genes larger than approximately one kilobase. Using gene amplification methods (e.g. polymerase chain reaction (PCR), Mullis *et al.*, U.S. Patent 4,683,195), together with synthetic oligonucleotides, it is possible to make 5 biologically active, synthetic retro-vectors that are capable of RNA transcription, reverse-transcription, viral packaging, and integration into genomic DNA (see for example, Hodgson, WO94/20608). Hodgson, *supra*, also disclosed methods for cloning of transcriptional promoters into such a vector using traditional recombinant DNA technology.

Modified restriction enzyme sites, linkers, and adapters can change the 10 primary or secondary structure of complex nucleic acid sequences thereby altering or obliterating a desired biological activity. For example, small mutations can drastically modify transcriptional promoters or change the reading frame of coding DNA. A logical goal of vectorology is to make exact constructs, without need of fortuitous restriction sites, adapters, or linkers.

15           Restriction endonucleases can be grouped based on similar characteristics. In general there are three major types or classes: I, II (including IIS) and III. Class I enzymes cuts at a somewhat random site from the enzyme recognition sites (see Old and Primrose, 1994. *Principles of Gene Manipulation*. Blackwell Sciences, Inc., Cambridge, MA, p.24). Most enzymes used in molecular biology are type II enzymes. These enzymes recognize a 20 particular target sequence (i.e., restriction endonuclease recognition site) and break the polynucleotide chains within or near to the recognition site. The type II recognition sequences are continuous or interrupted. Class IIS enzymes (i.e., type IIS enzymes) have asymmetric recognition sequences. Cleavage occurs at a distance from the recognition site.

These enzymes have been reviewed by Szybalski *et al.* *Gene* 100:13-26, 1991. Class 25 III restriction enzymes are rare and are not commonly used in molecular biology.

U.S. Patent No. 4,293,652 employed a linker with a class IIS enzyme recognition sequence to permit synthesized DNA to be inserted into a vector without disturbing a recognition sequence. Brousseau *et al.* (*Gene* 17:279-289, 1982) and Urdea *et al.* (*Proc. Natl. Acad. Sci. USA* 80:7461-7465, 1983) disclose the use of class IIS enzymes for 30 the production of vectors to produce recombinant insulin and epidermal growth factor respectively. Mandecki *et al.* described a method for making synthetic genes by cloning small oligonucleotides using a vector (*Gene* 68:101-107, 1988). Expansion of a population of

oligonucleotides required synthesis, cloning excision and fragment purification. The oligonucleotides were used to create a complete plasmid.

Lebedenko et al. (*Nucl. Acids Res.* 19(24):6757-6771) illustrated the class IIS enzymes and PCR for precisely joining 3 nucleic acid molecules for convention sub-cloning using BamHI. Tomic et al. (*Nucleic Acids Res.*, 18:1656, 1990), reported a method for site-directed mutagenesis using the polymerase chain reaction and class IIS enzymes to join two nucleic acid molecules. Two overlapping PCR primers were used where the primers included class IIS recognition sites. The primers included a region of complementarity to the template DNA and include one to a few site-directed mutations. Stemmer et al. (U.S. Patent No. 5,514,568) employed overlapping primers with class IIS enzymes to amplify a plasmid and to introduce specific mutations into DNA leaving all other positions unaltered.

There remains a need for the ordering and assembly of complex genes to overcome the problems associated with sequential sub-cloning such as multiple purification steps, the potential for sample loss, and the like. Moreover there is a need for eliminating the use of prokaryotic hosts and for minimizing or avoiding the risks associated with bacterial contamination resulting from the use of bacteria as intermediaries in the cloning process. Further, there remains a need for efficient methods to assemble large nucleic acid molecules or many-fragmented nucleic acid assemblies with precision.

#### 20 Brief Description of the Figures

**Fig. 1A.** provides one schematic of six double stranded DNA fragments, each terminus comprising a unique overhanging two-nucleotide sequence complementary to only one other terminus

25 **Fig. 1B.** illustrates a three-piece ligation where 100% of the clones tested contained the predicted fragment order and desired fragment orientation.

**Fig. 2.** illustrates the use of a class IIS restriction endonuclease (as one example, *Bpm1*), restriction endonuclease recognition site and the selection of cohesive overhanging ends.

30 **Fig. 3A.** illustrates an exemplary retrotransposon-derived vector including a murine VL30 LTR (NLV-3) and packaging signal, an internal ribosome entry site (IRES) from encephalomyocarditis virus (EMCV), a gene encoding a green fluorescent protein (GFP), additional internal VL30 sequences (solid bar), SV40 early region promoter and Tn5

aminoglycosidase phosphotransferase (neo) gene, PBR322 plasmid origin of replication and a plus-strand primer binding site (VL30). An exemplary vector sequence is provided as VLBPGN (SEQ ID NO:1). Fig 3B is an illustration of an LTR with the insertion of a U3 (transcriptional promoter)region rescued by reverse transcriptase-polymerase chain reaction (RT-PCR). The promoter is amplified from the RNA of a cell expressing the VL30 U3 region. Complementary overhanging ends are created using class IIS restriction endonuclease digestion sites within the LTR and within the promoter. Fig. 3C provides the linear structure of a VL30 RNA transcript from a mouse cell with a U3 region near the 3'-terminus of the RNA molecule. PCR primers include a class IIS enzyme recognition site to amplify the U3 region from the RNA resulting in a double stranded DNA molecule. Cleavage with a class IIS enzyme (here *BpmI*), results in a double-stranded DNA molecule with end complementary to a site in the vector of Fig. 3A.

Fig. 4A. is a schematic illustrating steps for assembling a combinatorial library of *cis*- or *trans*-acting nucleic acid sequences for assembly and screening, useful for the rescue of biologically active species. Fig. 4b is a diagram of a U3 (transcriptional enhancer and promoter region of an LTR illustrating several sub-divisions of the transcriptional control region, including a distal enhancer region, an enhancer repeat region, a medial promoter and a proximal promoter. These regions have been described for other vectors in Hodgson et al. (1996. "Construction, Transmission and Expression of Synthetic VL30 Vectors" in Hodgson ed. *Retro-vectors for Human Gene Therapy*. RG Landes Company, Austin TX). Segments of these regions are amplified using primers for highly conserved sequences. Highly conserved sequences are determine based on a comparison of known VL30 sequences such as provided in Fig. 4.2 of Hodgson, 1996, *infra*). The parts are joined by annealing and ligation to provide an ordered assembly. Each construct is an allele or a representative of allelic variation in the combinatorial library.

Fig. 5 discloses two transcriptional promoters that have been rescued from mouse VL30 RNA sequences isolated from a mouse T-helper cell library. These promoters were assembled into a vector and introduced into retroviral helper cells and packaged into recombinant retrovirus for introduction into human T-cells. After transduction to human T cells, a  $\beta$ -galactosidase reporter gene was expressed from the T cell-derived promoters.

Fig. 6 discloses 10 biologically active mouse VL30 promoters obtained from mouse liver RNA. These promoters were introduced into the vector of SEQ ID NO:1. The vectors

were introduced into retroviral helper cells and then packaged into retrovirus where they were introduced into human liver cells. The cells expressed the green fluorescent protein.

**Fig. 7** illustrates a similarity plot of nucleotide sequences found in VL30 U3 regions.

**Fig. 8** illustrates a retro-vector comprising six double-stranded DNA fragments that  
5 were self-assembled into a circular structure using unique overlapping termini created using class IIS restriction endonucleases. Three templates and twelve primers were used in conjunction with three class IIS enzymes to make the six fragments that were ligated in a single step. The vector was efficiently self-assmebled and was effectively transmitted by both DNA transfection as well as by retroviral transduction of the self-assembled DNA,  
10 without molecular cloning through a prokaryotic host (see Example 2).

#### BRIEF SUMMARY OF THE INVENTION

The invention described herein provides seamless, directional, ordered construction of complex DNA molecules, vectors and libraries. More particularly, it enables  
15 gene constructs to be assembled with greater efficiency and precision, and it enables multiple gene fragments to be assembled in the correct order and orientation without disturbing the internal structure of the gene. The method utilizes *in vitro* assembly of nucleic acid fragments and relies upon the unusual ability of certain enzymes to digest nucleic acid molecules at pre-determined sites without disrupting the structure of the gene. It is especially  
20 useful for the construction of genetic vectors for gene therapy or genetic engineering of cells and organisms. A particular application of the invention is in combinatorial, or evolutionary genetics, where it enables a large number of non-random, self-assembled constructs to be screened simultaneously for function.

In a preferred embodiment of this invention, the invention relates to a method  
25 method for assembling a gene or gene vector comprising the steps of: a) designing at least 6 primers to produce to amplify at least three fragments in at least three separate polymerase chain reactions wherein each primer comprises at least one predetermined restriction endonuclease recognition site that recognizes a restriction endonuclease that cleaves at a distance from the recognition site, a sequence complementary to a template nucleic acid for  
30 amplification, and bases positioned at the restriction endonuclease cleavage site that are selected to be complementary to only one other overhanging created from enzymatic cleavage of the fragments; b) combining the primers with template nucleic acid and performing the

polymerase chain reaction to produce multiple copies of an amplified template fragment incorporating the restriction endonuclease recognition site; c) digesting the amplified template fragments with one or more restriction endonucleases that recognize the restriction endonuclease recognition site of the primers to create overhanging termini wherein each overhanging termini is complementary to only one other overhanging termini on another fragment; and d) combining the amplified and digested template fragments in a ligation reaction to produce a directionally ordered gene, nucleic acid fragment or gene vector.

In a preferred aspect of this embodiment, the restriction endonuclease is at least one class IIS restriction endonuclease and preferably, the class IIS restriction endonuclease is selected from the group consisting of: *AlwI*, *Alw26I*, *BbsI*, *BbvI*, *BbvII*, *BpmI*, *BsmAI*, *BsmI*, *BsmBI*, *BspMI*, *BsrI*, *BsrDI*, *Eco57I*, *EarI*, *FokI*, *GsuI*, *HgaI*, *HphI*, *MboII*, *MnlI*, *PleI*, *SapI*, *SfaNI*, *TaqII*, *Tth111II*. Still more preferably, class II restriction endonuclease recognition sites (to be distinguished from class IIS restriction endonuclease recognition sites), linkers, or adapters are not used to create the gene or gene vector. In one embodiment, the product of the ligation reaction is introduced into prokaryotic or eukaryotic cells. Preferably, at least one template nucleic acid sequence is chosen from the group consisting of : transcriptional regulatory sequences; genetic vectors; introns and/or exons; viral encapsidation sequences; integration signals intended for introducing nucleic acid molecules into other nucleic acid molecules; retrotransposon(s); VL30 elements; or multiple allelic forms of a sequence.

In another preferred aspect of this embodiment, the method is used to generate combinatorial libraries of a target sequence. Preferably, the target sequence is part or all of a gene. In one embodiment, the gene encodes a protein. In one embodiment, the primers amplify allelic variants of part or all of a gene.

In still another preferred aspect of this embodiment, the product of the ligation reaction is passed between eukaryotic cells using a virus particle, by cell fusion, or by transfection. Preferably the product of the ligation reaction is not introduced into prokaryotic cells. Moreover, the method further comprises combining at least one screening or selection step to select the products of the ligation reaction. In one embodiment, the product of the ligation reaction is mutated during passage in cells in order to generate genetic diversity and preferably the product of the ligation reaction is mutated by homologous recombination during passage in cells.

In another aspect of this embodiment, the method is used to isolate and identify regulatory sequences from a cell. In another aspect of this embodiment, cells containing the product of the ligation reaction are selected for enhanced biological activity. Preferably, the cells containing the product of the ligation reaction are selected for tissue-specific, hormone-specific or developmental-specific gene expression. Also preferably, the ligation reaction is a circularized gene vector.

In another embodiment of this invention, the invention relates to a nucleic acid primer having a 5' and a 3' end to amplify a nucleic acid fragment for the ligation of at least two fragments comprising: a restriction endonuclease recognition site that recognizes a restriction endonuclease, wherein the restriction endonuclease cleaves at a distance from the recognition site and creates overhanging termini; a sequence complementary to a template sequence to be amplified to produce the nucleic acid fragment; at least two nucleic acid bases positioned at the restriction endonuclease cleavage site and that form an overhanging terminus after cleavage by the restriction endonuclease, wherein the at least two nucleic acid bases are selected to be complementary to only one other overhanging terminus on another fragment of the ligation; and an affinity handle on the 5' end of the primer. Preferably the primer further comprises an anchor to provide stability to the restriction enzyme at the restriction enzyme recognition site.

In yet another embodiment of this invention, the invention relates to a method for isolating and identifying promoters comprising the steps of: a) obtaining a vector comprising at least a portion of a promoter region from a retrovirus transposon LTR and having two non-complementary overhanging termini; b) designing at least two PCR primers to amplify at least one region of a retrovirus transposon LTR from template nucleic acid to produce at least one nucleic acid fragment wherein each primer comprises at least one predetermined restriction endonuclease recognition site that recognizes a restriction endonuclease that cleaves at a distance from the recognition site, a sequence complementary to a template sequence from a retrovirus transposon, and bases positioned at the restriction endonuclease cleavage site that are selected to be complementary to only one other overhanging terminus of the vector wherein the restriction endonuclease cleavage site is created from enzymatic cleavage of the fragments; b) combining the primers with template nucleic acid and performing a polymerase chain reaction to produce multiple copies of an amplified template fragment incorporating the restriction endonuclease recognition site; c)

digesting the amplified template fragments with one or more restriction endonuclease that recognize the restriction endonuclease recognition site of the primer to create overhanging termini; and combining the amplified and digested template fragment in a ligation reaction with the vector to produce a gene vector with an intact LTR sequence. In one embodiment of 5 this aspect of the invention, the template nucleic acid is DNA or RNA. In another embodiment of this aspect of the invention, the method further comprises the step of sequencing the insert to identify the promoter sequence. In one embodiment promoter sequences of SEQ ID NOS:1-13 identified using the methods of claim.

10

### Detailed Description of the Invention

In one embodiment of this invention, the invention relates to the seamless, oriented self-assembly of at least three DNA fragments having overlapping unique cohesive ends generated by the enzymatic cleavage of at least one restriction endonuclease that is capable of cleaving at a site distant to the restriction enzyme recognition site. Preferably the 15 restriction endonucleases employed in this invention are class IIS restriction endonucleases. These enzymes recognize a predetermined group of nucleotides and cleave at a distance characteristic of the particular endonuclease from the recognition site. The term "unique cohesive ends" is used herein to refer to the notion that the cleavage site for the 20 endonucleases of this invention can be manipulated to produce overhanging ends with unique termini selected by the investigator. The term "complementary" as used herein in reference to the overhanging ends of the fragments of this invention refers to standard complementarity recognized in the field of molecular biology. For example, the nucleotide sequence 5'-TAG- 25 3' is said to be complementary to the nucleotide sequence 5'-CTA-3'. The term "PCR" is used generally to refer to the polymerase chain reaction and its variations, including RT-PCR as well as other gene amplification techniques employing primers.

In a first step for practicing one embodiment of this invention, a series of at least three overlapping fragments are created through the selection and creation of primers incorporating at least one class IIS restriction enzyme recognition sequence. The oligonucleotide primers of this invention are designed to amplify one or more nucleic acid 30 fragments and comprise a sequence complementary to a target sequence for gene amplification, a recognition sequence for a restriction endonuclease that cleaves DNA at a distance from the recognition sequence (such as a class IIS restriction enzyme) and bases

positioned at the restriction endonuclease cleavage site that are preferably unique and complementary to only one other overhanging termini in the annealing/ligation reaction that generates the complex nucleic acid molecules. Optionally, the primers of this invention can include an "affinity handle for cleanup" at the 5'end. These sequences can be of any length, 5 preferably at least about 6 bp and the sequences extend the primer in the 5' direction from the restriction enzyme recognition site. This extra length gives many enzymes greater stability and improved activity. In addition, the sequence can be used for recognition and removal of the ends of the primers (either undigested fragments or digested ends of primers) using complementary nucleotide sequences bound to a solid support (such as cellulose, 10 nitrocellulose or silica). Incubation with, or passage over a column or support containing the complementary sequences can be used to remove the tags by allowing them to anneal or hybridize. The nucleic acid can then be eluted from the column. Adapters can also be used in this invention. For purposes of this invention, adapters refer to double stranded fragments containing an enzyme recognition site, according to this invention. The adapters are ligated 15 to double stranded DNA molecules, creating a fragment analogous to a PCR fragment with similar sites derived from a primer. The primers or adapters can be prepared using a number of methods for synthesizing oligonucleotides known in the art. For example instruments for producing oligonucleotides are available from Applied Biosystems, Inc., Foster City, CA.

In one example, for the design of an oligonucleotide primer for use in this 20 invention, the particular complementary bases that will form the site for hybridization of the primer to template (i.e., target DNA or RNA) are selected. A restriction endonuclease recognition site is selected followed by a number of nucleotides to be positioned between the recognition site and the cleavage site. The nucleotides of the cleavage site are selected to include overhanging regions formed from the restriction endonuclease cleavage that are 25 complementary to the overhanging regions of an adjacent fragment in the annealing/ligation reaction.

The length of the primer used in this invention can vary, but preferably the 30 primer length is up to about 80 bases and preferably up to about 50 bases. In addition the primers are preferably at least about 15 bases in length and preferably at least about 25 bases in length. The 5' region of the primer contains preferably at least about 6, preferably at least about 10 and still more preferably at least about 16-18 bases that are not complementary to the template DNA or RNA. Further, the primer incorporates a restriction endonuclease

recognition site preferably 5' to the region of complementarity and a restriction endonuclease digestion site preferably 5' to the region of complementarity or within the region of complementarity. There are a variety of restriction endonucleases that cleave at a distance from the restriction endonuclease recognition site of a DNA strand and a variety of enzymes 5 that are commercially available from New England Biolabs are provided in Table 1.

**Table 1. Restriction endonucleases useful in the construction of self-assembling genes**

Enzyme:	Site size (bp):	Distance to overlap:	Size of overlap:	Overlap type:
<i>Alw26 I</i>	5	1-5bp	4bp	5'-Overhang
<i>BbsI</i>	6	2-6bp	4bp	5'-overhang
<i>BpmI</i>	6	16-14bp	2bp	3'-overhang
<i>BsmBI</i>	6	1-5bp	4bp	5'-overhang
<i>BspMI</i>	6	4-8bp	4bp	5'-overhang
<i>BsrDI</i>	6	0-2bp	2bp	3'-overhang
<i>Eco57I</i>	6	16-14bp	2bp	3'-overhang
<i>FokI</i>	5	9-13bp	4bp	5'-overhang
<i>Hgal</i>	5	5-10bp	5bp	5'-overhang
<i>HphI</i>	5	8-7bp	1bp	3'-overhang
<i>MnII</i>	5	7-6bp	1bp	3'-overhang
<i>PleI</i>	5	4-5bp	1bp	5'-overhang
<i>SapI</i>	7	1-4bp	3bp	5'-overhang
<i>SfaNI</i>	5	5-9bp	4bp	5'-overhang

10 In addition to the enzymes provided in Table 1, other restriction endonucleases that cleave at a distance from their restriction endonuclease recognition site include, but are not limited to, *AlwI*, *BbsI*, *BbvI*, *BbvII*, *BsmAI*, *BsmI*, *BsrI*, *EarI*, *GsuI*, *MboII*, *TaqII*, *Tth111II* and their respective isoschizomers. These and other enzymes are known in the art and many are available from other manufacturers. The primers can be prepared to produce 15 either 5'-overlapping ends or 3'-overlapping ends, as long as they are both are either 5'-overlapping ends or 3'-overlapping ends and are complementary to one other set of overlapping ends.

In the case of *BpmI* (see Example 1), the enzyme digests asymmetrically, 14-16 bp from the 3'-nucleotide of the recognition site. The resulting cleavage has a 3'-overhanging end of 2 bp. A second primer is then designed with a complementary 20

overhanging end, and it is used to generate the adjoining fragment terminus. At the opposite ends of the two fragments that are to be joined, similar complementary, overhanging ends are designed.

The oligonucleotides are then combined with template nucleic acid (either 5 DNA or RNA, e.g., such as for reverse transcriptase polymerase chain reaction (RT-PCR)) containing bases complementary to at least a 3' portion of the primers (also referred to herein as "templates"). In one embodiment, the fragments are gene-amplified by PCR, RT-PCR or another gene amplification process using established PCR protocols such as those provided with PCR amplification kits, including those available from Perkin-Elmer Corp. (Emeryville, 10 California). Preferably, the PCR products are analyzed by electrophoresis on a gel, such as an agarose gel and still more preferably the fragments of the predicted size are purified free of excess primers and small byproducts (such as by purification through a small column, such as a Qiagen™ column (Qiagen, Valencia, CA)). Following amplification or purification, the fragments are digested with the restriction endonuclease recognizing the restriction 15 endonuclease recognition site in the primers. The digested fragments are then purified from the digested ends of the primers, preferably by preparative agarose gel electrophoresis. The fragments are combined, annealed and are ligated using standard hybridization and ligation conditions known for cloning (see Ausubel et al., *Current Protocols in Molecular Biology*, John Wiley & Sons, 1994).

Fig. 1A illustrates an example of a self-assembling gene construct (SEQ ID 20 NO:1) comprising six fragments, each having unique overhanging dinucleotide ends. In this example, the ends of the fragments prepared by the methods of this invention are constructed using primers that include *BpmI* restriction endonuclease recognition sites. It will be understood by those of ordinary skill in the art that one or more other restriction 25 endonucleases (such as those of Table 1) could similarly be used for the self-assembling product of Fig. 1A. In a preferred embodiment, the primers were created as described above and preferably the 3'ends of the primers are non-palindromic (i.e., non self-complementary) to prevent self-annealing of such fragments. Each fragment in this example preferably joins to only one other dinucleotide overhang in the annealing/ligation mixture, assuring ligation only 30 to the intended fragment partner. An advantage of this strategy is that the formation of concatamers or multimers is minimal. The restriction endonuclease site is removed by

digestion with the restriction endonuclease, leaving the junction free of the extra DNA sequences associated with the site.

Using a single restriction endonuclease with a dinucleotide overhang (for example, using the enzyme *BpmI*) up to six pieces of genetic material can be joined together 5 in a linear or circular form (such as a vector) without the need to perform sub-cloning procedures or detailed analysis of individual products because six unique combinations of dinucleotide overhangs create a directional clone with extremely high fidelity. With enzymes digesting single-base overlaps, only two fragments can be joined with positional and 10 directional precision. With enzymes digesting three-base overlaps,  $4^3/2$ , or 32 fragments can be so joined in the correct order and orientation. Therefore, this invention also relates to the use of restriction endonuclease recognition sites that facilitate cleavage by restriction endonucleases with three-base overlaps and self-assembly gene constructs including 32 15 fragments. Alternatively, a combination of restriction endonuclease recognition sites for use with a combination of restriction enzymes that create two-base or three-base overlaps can be used. Each enzyme has its characteristic limits to self-assembly imposed by the size of the overlap. For example, there are sixteen dinucleotides, therefore *BpmI* fragments (which have 20 two dinucleotide ends each) are limited to eight for the purpose of self-assembly; therefore in another embodiment of this invention an assembly comprising eight fragments is contemplated. However, four of the sixteen dinucleotides are palindromes. Use of these palindromic dinucleotides can create some infidelity in the annealing/ligation reaction. The 25 enzyme *HgaI* has a five base overlap, and there are 1,024 pentanucleotide combinations, permitting 512 fragments to be ligated together directionally and in order (no palindromes). The fragments to be joined at a particular place are designed to have their cut sites aligned, so that the overlapping region fits together. In some cases, the target sequences will contain 30 natural restriction endonuclease recognition sites for the enzyme that is being used, such as one or more internal *BpmI* sites. These sites have the potential to self-religate during vector or gene construction or they can be bypassed by using a substitute enzyme in the primers (for example, Eco 571 can substitute for *BpmI*). Alternatively, these sites can be removed by site-directed mutagenesis after consideration to the consequences of the mutagenized sequence to the gene or vector.

In addition to class IIS enzymes, class II restriction endonucleases can be used. These enzymes have intrinsic methylation activity that affects the outcome in either a

negative or a positive way, depending on the purpose for which it is used. In a preferred embodiment, the methylation activity of class II enzymes is ablated by mutation or by genetic engineering to convert the enzyme to an effective class IIS enzyme to expand the repertoire of useful enzymes for this invention.

5       In another aspect of this invention, the primer design and target fragment sequence selection can be automated (see Example 5) using a computer to assist in the selection of unique overhanging ends that have complementarity only to the overhanging end of an adjacent fragment.

Therefore, this invention permits high-fidelity annealing and ligation of six or  
10      more fragments with unique overhanging termini complementary to a single other overhanging termini. Any multitude of combinations can be created by combining the type of overhanging termini that can be created. Moreover, if one is willing to sacrifice the fidelity of the reaction, a variety of combinations can be used to anneal a variety of fragment numbers. In these cases, some selection may be necessary, such as size selection of the  
15      resulting fragment based on electrophoretic migration or restriction endonuclease profiling, both methods well known to those of ordinary skill in the art

It is also necessary to have a high per-step efficiency (e.g., each step in the process is performed with an efficiency of at least 80%) to effectively ligate large numbers of fragments without error. Where large numbers of fragments are used, the purity of the  
20      fragments becomes important. This means that for large numbers of fragments, the digested DNA fragments for annealing and ligation should be substantially pure. If undigested fragments, digested ends of primers, degraded or partially degraded molecules are present, they can decrease the purity and affect the fidelity of the product. Therefore, it is particularly desirable to ensure complete digestion of both ends of each fragment and to remove all of the  
25      digested ends from the fragments prior to including the fragments in an annealing and ligation reaction. The use of Qiagen columns for oligonucleotide removal prior to digestion is generally sufficient to permit efficient digestion of the fragments. Agarose gel isolation is desirable after digestion particularly where the product contains some fragments that do not appear to be full length. The use of an analytical gel before and after digestion helps in  
30      determining whether both oligonucleotide tags have been removed. The isolation of fragments from agarose gels preferably avoids the use of ultraviolet light and exposure of the

DNA to ethidium bromide is also preferably avoided. These methods can be avoided by running replicate lanes and staining only a portion of the gel.

The fragments and vector are then digested to yield fully complementary ends, and the fragments are preferably again purified, as described above (such as through a Qiagen column or by gel isolation). The purified fragments are ligated together in a test tube, under standard conditions, such as using bacteriophage T4 DNA ligase and ATP. Preferred ligations include at least 20 $\mu$ g/ml total DNA concentration in the ligation mix to favor intermolecular interactions, and an equimolar ratio of fragments to be ligated. Where a prokaryotic intermediary is used, the ligated assemblage is transformed into a bacterium, such as an *E. coli* host, and the colonies are: selected with a drug (such as an ampicillin, tetracycline, or kanamycin marker). The colonies can then be selected either by individually selecting colonies or growing a mass culture, such as where a vector library has been created. Restriction enzyme analysis can be used to determine the identity of individual constructs or to assess the validation of the combination of plasmids. The plasmids can then be grown up and used as needed.

In one embodiment of this invention, at least a portion of a vector is used as one of the fragments for the ligation of at least three fragments according to this invention. In one example, where a vector is used as one of the starting fragments, two restriction endonuclease recognition sites recognizing an enzyme that cleaves at a distance from the recognition site, such as at least one *BpmI* site, can also be introduced into the vector. This permits the vector to be digested with the restriction endonuclease to produce a product having ends complementary to two ends of the insert DNA fragments. The vector can be made by amplifying a plasmid or portion thereof using the primers of this invention. Thus, the vector can also be constructed to include a variety of restriction endonuclease recognition sites using a variety of restriction endonucleases, including a variety of class II restriction endonucleases. In some cases, the target fragments for amplification will contain natural restriction endonuclease recognition sites for the enzyme that is being used for the self-assembly, such as for example, a fragment that includes one or more internal *BpmI* sites. Care should be taken either to utilize the complementarity of the naturally occurring site to reform the fragment as it originally existed or to eliminate the restriction endonuclease recognition site using, for example, site-directed mutagenesis. Preferably, the restriction endonuclease recognition site is substituted for a different enzyme (in the case of *BpmI*,

substituting *Eco*57I or *Bsr*DI) that has an equivalent structure at its ends. Two or more fragments of insert or two or more fragments of vector with at least one insert are amplified using primers according to this invention.

The exemplary enzyme, *Bpm*I digests DNA 14-16 base pairs (bp) from the 3'- nucleotide of the recognition sequence (RS). Thus, by placing the RS exactly 14-16 bp from the desired dinucleotide cut site, the practitioner tags the dinucleotide for ligation with another dinucleotide that is exactly complementary to it. Such a complementary dinucleotide can be inserted by using the same enzyme and RS to make another fragment which fits the first exactly, as illustrated in Fig. 1. Because there are sixteen possible dinucleotide combinations (including twelve combinations that do not have palindromic ends), it is possible to create up to six fragments with unique dinucleotides, and it is also possible to join them all together in a predetermined order and orientation (Fig 1A). In addition, the palindromic sequences (such as AT, CG, TA, and GC) could also be used, although inefficiency and incorrect ligation will result from the self-complimentarity of these sequences. It is furthermore possible and desirable to have three or more fragments joined in this way, such that the construct is circular as in Fig. 1, comprising a vector that may be grown in a bacterial and/or eukaryotic host cell. If the genetic construct is to be used as a vector, the vector should be designed to include a proper origin of replication to enable it to replicate in a particular cell. For example, a prokaryotic origin of replication such as a coliform plasmid origin of replication enables circular DNAs to be propagated in *E. coli* host cells. It is desirable to have at least one selectable marker, such as a neomycin marker that enables recovery of the clone through a selection process. It is also desirable, but not essential, to have two or more selectable genetic elements, to permit dual selection. For example, if one of the fragments contains a prokaryotic plasmid origin of replication, and another fragment contains a selectable marker, then the two fragments are both selectable, since the construct will grow in prokaryotic cells in the presence of a selection drug (such as ampicillin) only when both fragments are present. Drug selection can be combined with the methods of directed self-assembly to assure a high percentage of correct products. Because of the unique complementarity of the fragments, each contributes a selectable element that leads to recovery of a high percentage of correct products.

For prokaryotic vector construction, at least one fragment should contain a prokaryotic origin of replication and one fragment should contain a drug resistance marker

gene. However, an advantage of the methods of this invention is that the construct can be introduced directly into eukaryotic cells. Here no plasmid origin of replication is necessary and no prokaryotic selectable marker or other prokaryotic nucleic acid sequence is necessary. In cases where the vector is subject to regulatory approval or where optimal gene function is necessary, it may be undesirable to include prokaryotic sequences, such as extraneous plasmids or expressed prokaryotic fragments particularly if the sequences contain immunostimulatory sites that can lead to activation of the intracellular immune system and inactivation of a gene product (see Krieg et al., *J. Lab. Clin. Med.*, 128:128-133, 1996) or to avoid risks of endotoxin contamination. Moreover, the use of self-assembled product, according to the methods of this invention saves labor and time involved in the screening process.

Thus, in a preferred embodiment of the invention, the nucleic acid fragments are self-assembled *in vitro*, and are transferred directly into eukaryotic cells, by transfection, injection, or other methods known in the art. In one embodiment the cells receiving the assembled product of this invention are helper cells for recombinant virus assembly (including, but not limited to retroviral helper cells for retroviral or retrotransposon vectors, adenovirus helper cells for adenovirus vectors or herpes simplex virus helper cells for herpes simplex vectors). Alternatively, the assembled product can be introduced into cells along with a helper virus or the assembled product can be introduced into target cells for direct expression. The assembled product can be a vector, a minichromosome vector, a portion of a chromosome, or the like. In the preferred case of a retroviral vector, the genes are first transfected into a first helper cell line (such as ecotropic helper cells, GP+E86 (Markowitz et al. *J. Virol.* 862:1120-1124, 1988). The retrovirus-containing supernatant from these cells is then filtered (0.45mm Nalgene filters) preferably 48-72 hours after transfection and the filtrate is transferred to a second complementation retroviral helper cell line (such as PA317 retroviral helper cells, Miller et al., *Mol. Cell. Biol.* 6:2895-2902, 1986). After an additional 48 h, the second helper cell line is selected with the marker drug (such as the drug G418 for the selectable neomycin (neo) marker gene), until only drug-resistant cells remain. These cells contain stably integrated vectors that can be used to repeatedly transduce human cells. Advantageously, in the case of adenovirus vectors or other large eukaryotic -derived vectors including eukaryotic virus-derived vectors, it may be impossible to propagate them in prokaryotic hosts. The gene self-assembly method of the instant invention provides an

alternative to *in vitro* recombination method of gene construction by permitting large constructs to be constructed.

One advantage of introducing the assembled product of this invention into a helper cell line to produce recombinant virus for the introduction of a gene or nucleic acid complex into a cell is that the assembled product will be auto-selected by the cells during the packaging process. Therefore, even where the overhanging termini have palindromic sequences, where there is more than one (but preferably less than four) unique complementary matches for a particular overhanging termini, or where concatamers have formed, only the correct or functional assembled products are expressed, transmitted, and assembled into virus. When the virus is then introduced into cells, the use of a reporter gene or another selectable marker provides yet a second layer of security for the selection of cells containing a properly assembled construct. For example, where a retrovirus helper cell line is used to produce a recombinant retrovirus containing the product of this invention (for retrovirus, RNA transcribed from the DNA product of the invention becomes packaged into the virus particle), a retrovirus-derived vector is transcribed as RNA and transmitted by packaging the RNA in a retrovirus particle. In order to be properly transmitted as a virus, the construct must be: 1) transcribed as RNA in a vector producer cell; 2) packaged into viral particles; 3) reverse transcribed into double-stranded DNA (in the recipient cell); and 4) integrated into the host chromosome. Each of these steps requires specific *cis*-acting sequences that must be correctly positioned within the vector. Thus, passage via retrovirus (or by other virus) is a means of auto-selection for the essential sequences.

In one application of the methods of this invention, the methods are used to rescue expressed sequences from RNA, or genomic sequences from cell DNA without disrupting the promoter sequences. Cellular transcriptional promoters are typically difficult to identify and isolate because they are generally not included in the RNA molecule and often extend over a considerable distance in a chromosome. One application of this invention relates to a promoter rescue technique that permits the entire promoter, or a fragment of a promoter to be isolated and cloned directly into an expression vector without disruption of the flanking sequences. Promoter rescue techniques are known and include WO 94/20608 to Hodgson.

In a preferred embodiment of the invention, transcriptional promoters are cloned in a transcriptionally active manner for the selection and identification of new and/or

of tissue or cell-specific promoters enabling them to be used, selected, or screened for activity directly. For example, Fig. 3 illustrates one example of the formation of a vector for the incorporation of promoter sequences and the ultimate identification of those sequences using an exemplary plasmid VLBPGN (SEQ ID NO:1) as provided in Example 3, with *Bpm1* sites located within the locus of a retrotransposon (VL30) long terminal repeat (LTR). These methods preserve the structure and functionality of transcription factor response elements. The characteristic secondary structure of the LTR RNA remains very similar to the original LTR from which the promoter was rescued, thus preserving the important features of the original RNA/DNA molecule. Those of ordinary skill in the art will recognize that any of a variety of primers can be used with a variety of vectors and that the constructs of Figs 2 and 3 are exemplary and not limiting.

Fig. 2 illustrates the primers used to amplify the promoter insert (identified at a and c in Fig. 2), and the insert region of the LTR (boxed), both of which can be digested at the same nucleotide position with *Bpm1*, to ensure a proper and seamless fit. In this example, after digestion of the vector, the two *Bpm1* sites leave non-complementary ends (a 3'-CC overhang on one end, and a 3'-GC overhang on the other). Thus, the ends will not efficiently anneal or ligate to one another. However, the complementary termini of the insert serves as linkage, enabling the plasmid to be completed by ligation.

In the example illustrated in Fig. 2, the terminus on the 3'-side (GC) is palindromic. Palindromic termini are self-complementary and can therefore ligate to themselves or to an identical terminus facing the opposite way (forming concatamers in the opposite direction). Despite the presence of palindromic termini and despite the potential for reduced fidelity in the self-assembling process, a large percentage of clones obtained by inserting promoter sequences into VLBPGN were assembled correctly (20/23). These levels are reduced somewhat when three or more fragments are combined for self-assembly, according to this invention and preferably, the use of palindromic termini are avoided when even numbers of nucleotides are exposed as overhanging termini because with even numbers of nucleotides there is an axis of symmetry. As noted above, where five base overhangs are used there are 1024 possible combinations of five nucleotides  $[(4)^5]$ , yet none of them is palindromic.

The vector of Fig. 3 is an example of a particular type of vector that is known as a retrotransposon vector. Retrotransposon vectors are described and reviewed in Hodgson

et al., 1996 *Retro-Vectors for Human Gene Therapy*. RG Landes Company, Austin TX, chapter 5 and see US Patent 5,354,674 to Hodgson. This type of vector is derived from a mouse cellular retro-transposon element that has no essential viral or cellular genes, and that has little sequence similarity to a retrovirus. However, this RNA (known as VL30 [virus-like, 5 30S]) has all the necessary *cis*-acting structural elements (such as LTRs and primer binding sites) required for efficient transmission by a type C murine or primate retrovirus. Thus, it is a parasite transmitted by retroviruses that is also expressed as a cellular RNA in most mouse cells and tissues. This RNA becomes packaged into retroviral particles when the mouse cells become infected by retrovirus. The retrovirus then transmits the VL30 (or a VL30 vector) to 10 the next infected cell (which can be a human cell). The RNA is then reverse transcribed and integrated into the DNA of the host cell.

Some advantages of VL30 vectors (over retrovirus-derived vectors) are: 1) lack of viral genes and other sequence homology that could lead to replication competent retrovirus (RCR); 2) ability to be expressed long-term *in vivo*; 3) a variety of LTR 15 transcriptional promoters that can be expressed in various tissues and under the influence of various hormones and other stimuli; and 4) the ability to express genes in a number of cell types that are targets of gene therapy. An additional advantage is that VL30 parts can be switched with those of classical retrovirus-derived vectors. For example, the LTR or packaging signal of VL30 can be used in place of the equivalent retroviral signal. The ability 20 to make mixed, or chimeric retro-vectors is a special application of gene self assembly technology.

Using a specific primer set, such as that shown in Fig. 2, or others, as taught in this invention, it is possible to amplify the U3 sequences expressed in the RNA of many different types of mouse cells. This is done using standard RNA isolation methods (Ausubel 25 et al., *supra*), coupled with extensive digestion with ribonuclease-free deoxyribonuclease, to eliminate residual DNA. Thus, to obtain a promoter that is expressed in the liver, one isolates RNA from liver and uses an RT-PCR procedure, such as those known in the art, with the primers to amplify the desired promoters. Fig. 6 illustrates liver RNA-derived promoters obtained using the methods of this invention. However, the promoters can also be derived by 30 conventional PCR from cDNA libraries (Fig. 5 illustrates T cell-derived promoters that were obtained in this manner). It is also possible to use the well-known hormonal and pharmacological inducibility of VL30 LTRs to find LTRs that are responsive to peptides,

hormones, and cytokines (for a table and description of VL30 pharmacologic responses (see Hodgson et al., 1996 *Retro-Vectors for Human Gene Therapy*. RG Landes Company, Austin TX, chapter 4, and Fig. 4.2). Examples of substances inducing various VL30 promoters to high levels include: epidermal growth factor, basic fibroblast growth factor, insulin, 5 erythropoietin, glucocorticoid hormones, activators of cyclic 3'-5'AMP, and others. To rescue promoters with pharmacological responsiveness, cells or animals stimulated with the desired pharmacological agent are subjected to the RT-PCR procedure and the resulting U3 regions are cloned into a vector, (such as the exemplary VLBPGN) and are tested for inducibility. Standard RNA blotting procedures can be used before isolating VL30 10 promoters, to determine whether a particular drug or hormone causes induction of VL30 RNA expression in a particular mouse cell or tissue. After the promoter has been rescued, the vector is transmitted via retrovirus to the target cell (possibly a human equivalent of the mouse cell from which the promoter was rescued). After selection with the drug G418 (400-700 µg/ml, for 7-10 days) to select against cells not containing the vector, the target cell 15 population is challenged with the pharmacological agent of choice. Reporter gene expression (in the example, GFP) or RNA expression, as determined by RNA blotting, can be used as an assay of gene inducibility by the agent (for exemplary gene expression methods, see Chakraborty et al., *Biochem. Biophys Res. Commun.* 209:677-683, 1995).

Using any specific primer set designed for use with VL30 retro-elements and 20 using total cellular RNA from a particular mouse cell type as a template for RT-PCR, (using commercially available kits and methods therein) candidate promoter elements can be amplified. This method is useful for the identification of mouse-derived promoters and in particular the method is useful for the identification of cell-type specific or tissue-specific promoters from a mouse and for the selection of these promoters and the identification of 25 tissue-specific or cell-specific promoters that function in human cells. Thus, these types of vectors and the methods for using these vectors permits the identification of promoters to permit controlled transcription of a foreign gene. The promoters, originally obtained from the mouse, can be used to effect tissue-specific or cell-specific expression in a human or animal liver cell such as a hepatocyte, or in a human blood cell such as a T-helper cell or in an 30 erythrocyte (red blood cell). Methods are disclosed in Example 2 for the screening and selection of the promoters from a library of amplified promoter sequences. Other methods are well known to those of ordinary skill in the art. The specificity of the selected promoter

can be assessed, for example, by introducing a selectable marker under the control of the test promoter in question and introducing this construct into various cells to assess the ability of the promoter to selectively regulate expression.

The amplified fragments represent U3 promoter regions from any RNA species expressed in the originating cells and their abundance will be in approximate proportion to the number of expressed copies of RNA in the original mixture. Example 3 illustrates one example using a mouse T-helper cell cDNA library to produce amplified fragments representing U3 regions expressed in T cells. The vectors were efficiently expressed as RNA and protein in PA317 helper cells, and were transmitted by retrovirus into human T-helper cells, where they were integrated and expressed as protein in the form of a  $\beta$ -galactosidase reporter gene, as visualized by X-gal staining. The products of this experiment are provided in Fig. 5 and as SEQ ID NOS: 2 and 3 from T-helper RNA. The products of another experiment are shown in Fig. 6 as SEQ ID NOS: 4-13 from mouse liver RNA (by RT-PCR).

Examination of the different U3 sequences isolated from T cells and from liver revealed several things. First, the T cell U3 sequences were related to each other, as were the liver sequences. However, the two types of U3 sequences were quite different between the two sources (T-cell, Figure 5 and liver, Figure 6). Specifically, the liver sequences (Figure 6) appeared to be a closely related group, differing mostly by single point mutations, some of which may affect transcription factor binding sites. Some of the polymorphic sites included: a phorbol ester response element (VLTRE); a Rel/NF $\kappa$ b binding region, and a possible glucocorticoid response element (GRE). Some of these polymorphisms are illustrated in Fig. 6. The T cell-derived sequences (Fig. 5, SEQ ID NO:2 and 3), on the other hand, differed significantly in length, with SEQ ID NO:3 missing more than 120 bases (compared with SEQ ID NO:2) including putative binding sites for retinoids (RAR/RXR) and several elements contained within the enhancer repeat region (including a cAMP response element (VLCRE, or CREB/jun binding site), and putative serum response element (SRE, CARG, and NF1/IL6). SEQ ID NO:3 represented one out of five clones sequenced, while SEQ ID NO:2 represented four out of five. Possible sites of interactions between transcription factors and DNA can be observed by comparing the experimentally derived U3 sequences with those in Hodgson et al. ,(Retro-Vectors for Human Gene Therapy, 1996 Fig. 4.2 *supra*). In addition

to the deleted sequences of SEQ ID NO:2, there are a number of single base differences within the conserved regions of the two T cell-derived sequences.

Advantageously, a number of new VL30 promoter sequences (SEQ ID NOS: 2-13, *supra*) were identified using these methods despite the fact that VL30 RNA comprises only about 0.3% of cell mRNA represented in a cDNA library. Moreover, in each case, the cloned insert was isolated without the need to use linkers, adapters, or multiple cloning sequences such as those that are typically used for other library construction methods. The promoter sequences can be used in the vectors disclosed here to express inserted foreign genes or the promoter sequences can be substituted into other retroviral vectors, such as MoMLV-derived vectors or other VL30-derived vectors. Further, vectors containing the promoter sequences can be propagated in retroviral helper cells, such as PA317 (U.S. Patent 4,861,719 to Miller) or introduced into cells by chemical or physical transfection.

In another application of the methods of this invention, libraries of amplified sequences can be incorporated into vectors using two or more fragments and using the restriction endonucleases cleaving at a distance from their recognition sites. Preferably the vectors are created using six or more fragments and preferably greater than 10 or more fragments. For example, as applied to VL30 promoter sequences, because there are over a hundred VL30 retro-elements in the mouse genome, it is possible to amplify all of the promoter sequences *en masse*, and propagate them *en masse*, enabling screening by serial passage through helper cells (such as the PA317 helper cell line) or by means of a replication competent retrovirus, as illustrated in Examples 3 and 4. Conversely, the promoter region may be broken down into several sub-domains and permutations of each could be combined and screened to enhance the chances of generating a superior construct (Fig. 4B).

As an example of breaking a promoter region down into several sub-domains, Fig. 7 illustrates a similarity plot of nucleotide sequences found in VL30 U3 regions. Plot similarity was performed using the Plot Similarity program (Wisconsin Sequence Analysis Package, release 8.1, Genetics Computer Group, Madison, WI). This program plots the running average of the similarity among the sequences in a multiple sequence alignment. The sequences compared were those found in Fig. 4.2 of Hodgson, 1996, chapter 4 (*infra*). That is, the plot discloses the degree of conservation of VL30 promoter sequences among known VL30 promoters. From the figure, it can be seen that conserved sequences (close to 100% conserved) can be used as primer binding sites to amplify the adjacent sequences by PCR.

An allelic mixture of three fragment sets is then created to make a combinatorial library of promoters that can be positively selected, such as by using retroviral amplification of the active sequences. This, used in combination with the Fig. 4.2 (Hodgson, 1996, chapter 4 *supra*) can be used to determine regions of high similarity. Regions of high similarity within 5 the U3 region can be replaced with one another. Therefore, a library of permutations of these sections can be made by combining allelic pools obtained by amplifying the sequences from individual subsections, followed by ligating the subsections in the correct order using the methods of the instant invention for gene self-assembly. For example, sub-section 1 can include the distal enhancer (from the LTR 5'-end to the site of insert primer 2, see for 10 example the region defined by the insert primers 1 and 2 (SEQ ID NOS 55 and 56 of Example 4). In this way, using a plot similarity (such as Fig. 7), within each sub-section, the primers position fragments within a region of nearly 100% identity. Degenerate primers can also be used in these experiments to account for multiple nucleic acid base combinations 15 along a particular sequence. In each case, the primers preferably are designed to have a melting temperature that is compatible with the RT-PCR conditions being used, and the conditions should be those recommended by the manufacturer (preferably Perkin Elmer Corp., Emeryville, CA). In Example 4, a set of primers is given that can be used to amplify different U3 subsections, together with directions for assembling a combinatorial library.

It will be appreciated by persons of ordinary skill in the art that the methods 20 of the instant invention can thus be used to make allelic libraries of a variety of genes. For example, different allelic portions of a gene can be combined in a predetermined order and orientation to produce combinatorial libraries, without the need for fortuitous restriction sites separating the parts in the original construct, and without perturbing the important sequences joining the parts using the methods of this invention.

In this invention primers are constructed as described above. However, for the 25 generation of allelic libraries or more complex library constructs it may be helpful to include 5' tags into the 5' end of the primer. The purposes of the tag sequence are: 1) to provide extra nucleotides on both sides of the restriction endonuclease recognition sites (for more efficient digestion); and 2) to enable recovery of sequence tags or undigested fragments by means of 30 an affinity reagent (such as silica, magnetic beads, or nitro-cellulose containing the complementary sequences) for purification. The use of an affinity reagent permits the digested ends to be purified away from the digested fragments. Furthermore, if any

undigested ends remain after thorough digestion, the affinity reagent will remove them, further aiding in the purification. In one embodiment, affinity purification of the digested fragments is used in place of gel isolation, eliminating possible damage caused by ultraviolet light as well as possible damage caused by dye (e.g., ethidium bromide) binding to the DNA.

5 It will also be appreciated that a number of other variations to the primer sequences can be employed. For example, as discussed above, the enzyme recognition site for an enzyme that digests outside of its recognition sequence is included in the primer, so that the DNA digest creates an overlapping end that is complementary to one other terminus to which it will be joined. The enzyme recognition site can be moved to any location within  
10 the primer so as to digest the DNA at the exact location desired. The primer can also be programmed with a novel enzyme recognition sequence to add any desired sequences between the two sequences to be joined or to incorporate a linker or adapter if desired. If the sequences to be amplified contain the enzyme recognition site of the primers, it may be necessary to switch to a different enzyme usage. The use of several different enzymes is  
15 possible and has been discussed above. As with other PCR procedures, after the initial primer selections have been made the primers are assessed for their ability to fold back on themselves or to create internal secondary structure. The primers are preferably modified to avoid palindromic sequences or the potential for self folding within a primer. Nucleic acid analytical software (such as the Wisconsin GCG package, Oxford Biomolecular, Oxford, UK)  
20 is available to perform this analysis and aid in the selection of alternative primers.

In addition, as with all PCR processes, it is necessary to determine the melting temperatures ( $T_m$ ), and to adjust the annealing temperature of the PCR reactions to compensate for such temperatures. Finally, it is important to perform a sequence redundancy search, to determine whether the target sequence (the sequence complementary to the primer)  
25 is found more than once in the region to be amplified. If the sequence is repeated, it will be necessary to use a different primer in order to establish the single, correct priming site. Preferably, no more than 6-8 bases of incorrect target complementarity at the 3'-end of the complementary region is used and to allow a difference of at least 10° C between the  $T_m$ s of the correct and the incorrect target. The annealing temperature should always be at least 5°C  
30 lower than the  $T_m$  of the correct target and 5°C above the  $T_m$  of the incorrect target. Again, the necessary software and instructions are readily available from the cited sources (Wisconsin Gene Computer Group and Oxford Biomolecular, *supra*)

Next, a vector is constructed to include the appropriate elements for expression in the desired cell type. For example, the plasmid of Fig. 3A can be used for the creation of a promoter library or a vector can be created using a commercially available vector and primers to create a three or more fragment annealing and ligation reaction as provided above.

- 5 Preferably, the inclusion of a dominant negative selectable marker on the vector (e.g., the neomycin phosphotransferase gene, conferring G418 drug resistance) can be used to reduce the likelihood that cells without the vector are being maintained in culture.

Multiple allelic copies of DNA (cell derived or cDNA) can be amplified in separate reactions as a set of potential inserts with each set having its own unique overlap 10 sequence following digestion with a restriction endonuclease, according to this invention. The fragments can then be ligated into an existing vector or in a single reaction of three or more fragments to form a combinatorial collection of potential alleles. For example, if six adjacent regions are amplified from five separate alleles, the number of combinations would be  $5^6$ , or 15,625 potential combinations. The combinations can then be grown *en masse*, and 15 selected *in vitro* or *in vivo*. A variety of screening strategies can be used in this invention and those of ordinary skill in the art will appreciate that the type of screen will match the type of library being generation. Therefore, for the promoter library, introducing members of the library into particular cell types to assess for expression in one or more cell types versus the absence of expression in another cell type is evidence of tissue-specific or cell-specific 20 expression. For screening purposes, the libraries of this invention function like other libraries created through other methods. A variety of screening methods for a variety of libraries have been described in the art. For example, selective screens are reviewed by Hodgson et al. (1996, RG Landes Company, *supra*). Reporter protein production is well known in the art as 25 is dominant selectable marker (e.g. drug) selection and selection by fluorescence activated cell sorting, antibody affinity selection, phage display selection (such as commercially available from Amersham, Milwaukee, WI), and the like can be used without detracting from this invention.

In this way, it is possible to isolate multiple forms of genes, gene fragments or regulatory regions such as transcriptional promoters or packaging signals (for example, in a 30 retro-vector system). The individual constructs may then be tested *in vitro* or *in vivo* to further characterize a particular phenotype.

In one example the method is used to create a library of complementarity determining regions (e.g., allelic variations that give rise to antibody diversity) of antibodies or from receptors, including T-cell receptors, epitopes, antigens, ligands and the like. For example, where a library of T-cell receptors is created, the introduction of a vector designed 5 to create a functioning T-cell receptor can be introduced into T cells or T-cell progenitors and the cells can be tested for their ability to bind to a particular test ligand. The ligand-recognizing cells can then be isolated from the ligand and grown in the presence of cytokines to produce specialized T cell clones. Where a library of antibodies or antibody fragments is created, the antigen reactive portions, for example, can be recombined in a vector containing 10 the remaining portions of an antibody molecule to generate antibodies or antibody fragments in a cell. In other examples, the methods of this invention can be used to create allelic domains of receptor families (such as the steroid receptor super-family); libraries with related regions from peptide hormones; cytochromes P450; or other protein families that have shared domains or sub-sections with similar structures. The methods of the instant invention allow 15 the joining of allelic sub-sections in an ordered fashion. In each case, it will be necessary to design primers, and to keep track of the uniqueness of joining overlaps and the presence of internal restriction sites as described above. While these will be different in each case, here are listed some general guidelines that are incorporated into the method of the instant invention.

20 As discussed above, although described as it relates to promoter libraries, libraries of other nucleic acid sequences can be created using the methods of this invention. These libraries include, introns and/or exons and/or functional domains libraries, libraries of potential alleles for a particular gene sequence, and the like. These sequences can be amplified from cell DNA or RNA using the primers of this invention and incorporated into a 25 variety of vectors. For example, one vector of this invention, VLBPGN, has a portion of LTR removed and can be used to create a variety of libraries following digestion with *Bpm*1.

Selected or screened products of the combinatorial library can be used for gene expression, such as the promoters of Figs. 5 and 6. In addition, the exploitation of these sequences for the expression of a variety of genes, the LTR fragment containing the promoter 30 can be joined to one or more functional retroviral packaging signals, internal ribosome entry sites, additional promoters, coding regions, processing sites, and the like.

Advantageously, there are almost no spatial constraints upon the joining of molecules by the method of the instant invention and other methods have not taken advantage of the combination of PCR to isolate genes or gene fragments; enzymes cleaving at a site distant from their restriction endonuclease recognition site to combine three or more fragments with precision; and, the use of unique overlapping non-palindromic termini to ensure fidelity of multi-fragment ligations. This combination permits the artisan to prepare complex gene constructions in one ligation step and does not require sequential sub-cloning into a vector or propagation in a prokaryotic host. Added to this the combination by these methods of fragment pools facilitates recombinatorial genetics.

The ability to recombine (in the correct order and direction) and screen a large number of allelic variants (whether as a simple library or as a combinatorial library), resulting in increased abundance (by amplification in the RNA, and subsequently in the DNA) is a special characteristic of this invention. Particular advantages of this system are obtained when the methods of this invention are combined with retrovirus vector technology or other virus vector technology. For example, the combination provides a form of *in vitro* evolution whereby the passage of the library through virus and through cells selects functioning sequences and increases the abundance of the surviving RNA and DNA molecules.

For example, consider the consequences of screening several different promoters expressing RNA in a donor cell (*i.e.*, a cell producing virus particles), but at differing levels of RNA abundance. In the following example, the least abundant RNA species is expressed at 0.1 copy of RNA per cell, while six others are expressed at 1 copy, 10 copies, 100 copies 1,000 copies, or 10,000 copies, or 100,000 copies/cell, respectively. After a single passage, the DNA copy number in the recipient cells now reflects the approximate RNA copy number in the donor cells. These numbers are further amplified in the relative abundance of RNA species produced in the recipient cells. Disallowing for factors such as position effects, transcription factor depletion, etc., (which may be considerable), the same relative ratios of expression would be expected. Taking into consideration position effects, the disparity between abundance caused by changing insertion loci should average out. The most abundant RNA species after two passages is then many orders of magnitude more abundant than the least abundant.

<b>Species:</b>	<b>RNA abundance:</b>	<b>DNA copy no.</b>	<b>RNA abun.</b>	<b>DNA copy no.</b>	<b>RNA abun.</b>
	P=0	P=1	P=1	P=2	P=2
<b>A</b>	0.1 copy/cell	0.1	0.01	0.01	0.001
<b>B</b>	1	1	1	1	1
<b>C</b>	10	10	100	100	1,000
<b>D</b>	100	100	10,000	10,000	$10^6$
<b>E</b>	1,000	1,000	$10^6$	$10^6$	$10^9$
<b>F</b>	10,000	10,000	$10^8$	$10^8$	$10^{12}$
<b>G</b>	100,000	100,000	$10^{10}$	$10^{10}$	$10^{15}$

**Table 2.** Enhancement of DNA and RNA copy number as a result of different RNA expression levels, after retroviral passage. P= (no. of passages). Numbers are interpreted as relative ratios within a column.

5

The present invention is able to efficiently create a library of RNA or DNA sequences whether or not they are in low abundance. The kinetics of screening for RNA abundance of a promoter can be appreciated best in the following discussion. For the purposes of this discussion, position effects have been ignored. An equation describing the 10 kinetics of screening for RNA abundance is:

$$(1) R_{\text{rel}\chi} = A_\chi / \sum A_{\alpha\infty}$$

The above equation (1) can be stated in plain English: The relative abundance 15 of an RNA species  $\chi$  ( $[R_{\text{rel}\chi}]$ ) within a population of RNA molecules expressed in a single cell or within a population of cells) is equal to the RNA copy number of RNA species  $\chi$  ( $A_\chi$ ) divided by the sum of the RNA copies of all RNA species present, including  $\chi$ .

The relative abundance number of any given species changes as the number of 20 passages change, according to the following approximation:

20

$$(2) R_{\chi\text{py}} = D_{\chi p0} R^{p+1}$$

In the simplest of terms, equation two (2) can be expressed as: The abundance 25 of RNA species  $\chi$  after Y passages ( $R_{\chi\text{py}}$ ) is equal to the initial abundance of the DNA for species  $\chi$  at passage=0 ( $D_{\chi p0}$ ), multiplied by the RNA abundance/DNA copy, raised to the power of the number of passages plus one. Thus, a typical RNA species that starts out as a

single copy of DNA, after zero passages (*i.e.*, in the donor cell) expresses 10 copies of RNA/cell. After one passage it is amplified at the DNA level to a relative ten copies (the same as the RNA abundance at P=0), and at the RNA level to 100 copies (10 copies per DNA copy). The reason for the amplification is that viral packaging and passage is based upon the 5 number of RNA copies present in the donor cell. These calculations can be used to arrive at approximate abundance determinations for any given passage. The actual results of any given experiment, of course, will be biological rather than physical or mathematical. This means that other variables such as RNA efficiency of transmission and longevity, availability of transcription factors, experimental variation, *etc.* also come into play. The underlying 10 purpose of the approximating equations, however, is to illustrate that RNA is amplified in DNA in proportion to the abundance of the template (RNA) within the cell.

The abundance of mRNA in cells can vary continuously from less than a copy per cell to nearly 100,000 copies/cell in actively transcribing, highly-specialized cells such as reticulocytes, the chicken oviduct, the silk moth silk gland, etc. Therefore, the spectrum of 15 RNA abundance from 0- $10^5$ /cell is within the biological window of interest. For most practical purposes, such as biotechnological expression of genes in specific cells, only the higher end of this abundance range is desired. Therefore, using a viral selection system, as disclosed in this invention, it may be possible to disregard those species with less than a threshold level, such as <0.1 copies per cell. The selection through virus will lead to the 20 recovery of the more abundant species. Furthermore, because the vector is likely to be the only considered sequence, it may be considered as a proportion of the whole of RNAs expressed in the target cell. The situation is more complex when a large number of permutations and combinations is generated, for example by self-assembling thousands or millions of fragments in a predetermined order using the self-assembly technique of the 25 instant invention. Consider the assembly of allelic variants of four promoter subregions: distal enhancer, proximal enhancer, distal promoter and proximal promoter. If 100 varieties of each of the four groups were amplified and combined using the instant process along with a single vector,  $10^8$  resultant combinations could occur. However, a sufficient number of molecules to start out a combinatorial screening program might be a million. The problem 30 can be simplified by considering these in groups as follows:

**Table 3. Grouped abundance of RNA molecules derived from combinations.**

No. of species in group:	RNA abundance:	Total No. RNA molec. at P=0:	RNA at P=1	RNA at P=2	RNA at P=3
9 X 10 <sup>5</sup>	1	9 X 10 <sup>5</sup>	9 X 10 <sup>5</sup>	9 X 10 <sup>5</sup>	9 X 10 <sup>5</sup>
2 X 10 <sup>5</sup>	10	2 X 10 <sup>6</sup>	2 X 10 <sup>7</sup>	2 X 10 <sup>8</sup>	2 X 10 <sup>9</sup>
2 X 10 <sup>4</sup>	1,00	2 X 10 <sup>6</sup>	2 X 10 <sup>8</sup>	2 X 10 <sup>10</sup>	2 X 10 <sup>12</sup>
1 X 10 <sup>3</sup>	1000	1 X 10 <sup>6</sup>	1 X 10 <sup>9</sup>	2 X 10 <sup>12</sup>	2 X 10 <sup>15</sup>
1 X 10 <sup>1</sup>	10,000	1 X 10 <sup>5</sup>	1 X 10 <sup>9</sup>	1 X 10 <sup>13</sup>	1 X 10 <sup>17</sup>
1	100,000	1 X 10 <sup>5</sup>	1 X 10 <sup>10</sup>	1 X 10 <sup>15</sup>	1 X 10 <sup>20</sup>
<b>Sum Total:</b>		<b>6.6 X 10<sup>6</sup></b>	<b>1.11 X 10<sup>10</sup></b>	<b>1.01 X 10<sup>15</sup></b>	<b>1 X 10<sup>20</sup></b>

Thus, it follows that in the example population (Table 3) of over a million

constructs (equally represented in the DNA), a single construct expressing 10<sup>5</sup> copies of RNA per DNA copy will increase to approximately 99% of the total expressed RNA sequences in

5 two passages. Using similar procedures in combination with drug and/or hormonal stimulation, and after consideration of the possible transcription factor binding sites within the sequence family (Figs. 5 & 6), it is within the intended scope of the invention to select for hormonal or pharmacological controls of transcription such as have been described herein. The factors contributing to the outcome are not only the input constructs, but recombinants 10 and mutants as well. These secondary contributors to molecular diversity will be enhanced if multiple rounds of infections are allowed to occur, as oftentimes the difference between a particular transcription factor being able to bind (or not) may depend upon a single base change. Because viral infection is progressive and competitive, molecular evolution can be used to generate gene constructs *de novo* in the tissue culture dish in short time periods.

15 Advantageously, the use primers to generate amplified fragments with uniquely complementary cohesive ends (i.e., that the ends will preferably only hybridize with the intended 5' and 3' fragments) to ligate three or more fragments as taught in this invention improves the potential for obtaining a diverse library.

Although the examples particularly point out a transcriptional promoter as the 20 product of the process, the skilled artisan can appreciate that a particular selection technique can be applied to other *cis*- and *trans*-acting genetic sequences as well. Although a virus is used to propagate the selective advantage of a preferred embodiment, it can also be appreciated that any selective screen, such as drug selection, cell survival, phenotypic selection, cell sorting, antibody selection, and the like (see Ausuble et al., *supra*) could be

substituted without changing the intended scope of the invention. Likewise, transfection or cell fusion could be used in place of viral infection. Furthermore, substitution of different viruses, retrotransposons, or functional groups are likewise within the intended scope of the invention. The described embodiments are to be considered only as illustrative and not restrictive, and the scope of the invention is indicated by the claims rather than by the narrative description. All references and publications, cited herein, are incorporated by reference into this disclosure.

Like the embodiments detailed above, the method of library production is also conducive to assembly and transfer of genetic material directly into eukaryotic cells, saving the step of propagation in bacteria that is standard in bacteria. An advantage of direct transfer of the libraries of this invention to eukaryotic cells, including the exemplary retroviral vector producer cells, is that certain essential *cis*-acting structural features will be under positive selection (i.e., if they are not present, the molecule will be lost due to its non-functionality). As discussed above, it is often advantageous to eliminate bacterial and plasmid DNA sequences, endotoxin, and other bacterial contaminants by introducing the constructs directly into eukaryotic cells.

In addition to providing a method for constructing complex DNA molecules efficiently (as in the examples of three piece and six piece constructs), the methods of this invention permit the assembly of constructs that are larger than those conventionally propagated in *E. coli*. Examples of these types of vectors include adenovirus vectors, herpes simplex vectors and artificial minichromosomes. In order to insert genes into such vectors that are too large for conventional molecular cloning procedures, in the past it was often necessary to resort to *in vivo* recombination, wherein the genes of interest are cloned into a suitable vector and the flanking homologous regions are used to target the foreign genes to a homologous site within the larger viral or minichromosome vector. However, the methods of this invention permit PCR fragments of any size (up to the limits of PCR capability, 20-30 kb per fragment) to be joined together. Thus, it is feasible to precisely construct adenovirus vectors by amplifying larger sequences, and combining them by ligation. For example, several sections of adenovirus (5-10 kb each) can be ligated using the methods of this invention, up to for example, about 37 kb, and then transformed directly into human cells. Only the correctly recombined vectors are capable of replicating. Hence, the DNA is autoselecting. A similar procedure is used for generating herpes virus vectors, which are

approximately 150 kb. The precision of the methods of this invention permit non-essential viral genes to be more easily eliminated from the construct. After transfection into appropriate cells, the DNA replicates and virus particles are formed.

Some special considerations apply to larger vectors, however. First, it is

5 desirable to use enzymes that do not cut within the large DNA fragments. To prevent excessive fragmentation of the DNA by internal sites, it is desirable to use enzymes that cut rarely or infrequently, such as CpG-containing enzymes recognizing six bases, or enzymes such as *Sap1*, recognizing seven bases and digesting a three bp overhang (thus permitting up to 32 fragments to be joined in order). It is also desirable to avoid shearing the DNA once  
10 large segments have been joined by ligation. One method of avoiding shear is to add the transfection agent, such as Superfect™ reagent (dendrimers, Qiagen) or Lipofectamine™ (liposomes, Life Technologies, Gaithersburg, MD) directly to the ligation reaction, and then add the cells to be transfected to the mixture. This, or a similar method avoids the need to physically move the ligated DNA, and thus prevents shearing. Another method is to add a  
15 DNA condensing reagent (dendrimers, polycations [such as polyethyleneamine] histones or liposomes) directly to the DNA ligation reaction, and then move the DNA by pipette after it has condensed (thus reducing shearing of the DNA). Once inside the cell, viral DNA can replicate (as in the examples of partially replication-competent adenovirus and herpes simplex virus vectors).

20 Artificial minichromosomes have been under development for years. True artificial chromosomes require a centromere, at least one origin of DNA replication, and in the case of linear molecules, telomeric repeats at the chromosomal termini. In addition, to be very effective it is desirable to have a selectable marker gene, one or more therapeutic genes, and/or reporter genes.

25 In reality, the use of minichromosomes has been delayed by the inability to effectively manipulate the larger DNA molecules *in vitro*. Yeast and bacterial artificial chromosomes have been used with little success in mammals, and the addition of telomeres to the ends of linear chromosomes is also a special problem, as there is no prokaryotic host that can tolerate large linear DNA. The methods of this invention offers the opportunity to  
30 assemble human or mammalian minichromosomes *in vitro*, by using large segments (10-30 kb) of synthetic, gene-amplified DNA as ligation starting materials. For example, up to 32 *Sap1* fragments (up to 30 kb each, containing the essential *cis*- and *trans*-acting sequences),

or 512 shorter *Hga*1 fragments can be combined using these methods. As with the other examples, several enzymes suitable for this invention (e.g., such as class IIS enzymes) can be combined (possibly with different termini lengths) to simplify the task. The methods of this invention also facilitate construction of telomeric repeats, because the constructs of this invention do not need to be circular. Thus, the methods of this invention can be used to make telomeres of any length, by adding additional segments onto the ends of molecules. One way to do this is using self assembling genes that employ a repeating overhang sequence (self-complementary molecule, such as AG-3' at one end, and CT-3' at the other end), permitting the telomeres to be lengthened to the extent desired by adding the required molar excess of the telomeric repeat-containing fragment. This technique gives the investigator some control over the relative length of the telomeres, although the self-complementarity indicates that many repeats will be lost due to self-ligation. This can be alleviated by using higher starting concentrations of DNA to favor inter-molecular ligations over intra-molecular ligations (e.g., >20 µg/ml starting concentration of DNA).

A two fold molar excess of telomeric fragments gives approximately twice the average length of telomere as a strictly 1:1 molar ratio of all fragments. By using a higher molar ratio of shorter telomeric repeats it is possible to give greater uniformity to the overall length of the molecules, which will vary from one terminus to the other. Thus, in addition to providing a way to build large molecules with precision, the methods of this invention provides for a way to control the telomere length (or potential life-span) of the artificial chromosome. To prevent damage during handling, the minichromosome DNA can be condensed with polycations, adenovirus particles, dendrimers, histones, or liposomes prior to transfection, similar to larger viral vectors.

The methods of this invention can be used to create recombinant virus. One example of this is an adenovirus vector self-assembling gene system. This system can include three parts: 1) vector; 2) helper virus; and 3) helper cells. The vector part is a self-assembling fragment set of at least three fragments comprising the essential cis-acting sequences (left and right inverted terminal repeats, which are the 103 bp at both ends of the genome that are required for replication [ITRs] and packaging sequences [Y, base pairs 194-358]) and central 'baggage' area, comprising one or more self-assembling fragments including therapeutic genes, marker genes, and reporter genes. The baggage area is thus flanked by the cis-acting sequences in the vector. Because the synthetic oligonucleotide sequences

comprising the 5' and 3' termini of the helper virus are not phosphorylated, they will not ligate together creating multimers. Thus, the Ad5 vector region will assemble only into monomers. The helper virus part comprises all Ad5 trans-acting genes except for the E1A and E1B genes. The helper virus part has no cis-acting sequences, and it is amplified in several sections. In this preferred embodiment, the virus is amplified using primers that exclude the ITRs, packaging region and E1A&B genes. The helper virus is digested by *Sap1* digestion, creating seven uniquely terminated fragments comprising the trans-acting viral genome, with dephosphorylated, blunt 5' and 3' ends on the terminating fragments. The primers are designed so as to amplify the internal virus sequences without changing them, except for the 5' and 3' ends of the virus. The PCR-amplified fragments are digested with *Sap1* and are religated in their natural order after gel isolation and Qiagen column purification. The 5' end of the helper virus genome starts at 3.2 kb (in the E1A gene) so as not to overlap the vector sequences, which could otherwise cause replication competent adenovirus (RCA). Because the 5' and 3' ends of the helper virus do not contain *Sap1* sites, they remain intact after digestion with *Sap1*. Because the synthetic oligonucleotide sequences comprising the 5' and 3' termini of the helper virus are not phosphorylated, they will not ligate. Thus, the Ad5 helper virus genome assembles only into preferred monomers during ligation.

In a preferred embodiment, non-essential genes are deleted from the Ad5 genome by means of the method of self-assembling genes. In another preferred embodiment, the helper virus genome is approximately 30 kb after deletion of E1A, E1B and E3 gene sequences from the helper virus, and it is amplified as a single long fragment using the eLONGase Amplification System (Life Technologies or a similar strategy for creating long PCR fragments with high fidelity). It is not of great importance that occasional PCR errors may occur, because multiple copies of the Ad5 helper virus are transfected into target cells, thus providing trans-complementation. The helper cells are preferably 293 cells, a human kidney cell line expressing E1A and E1B genes (ATCC). The vector part and the helper virus part are combined in equimolar ratios after ligation has been performed separately on each fragment set. The Superfect protocol (Qiagen) is used to transfet the vector part and the helper part into the helper cells. The helper cells lyse, releasing high-titer adenovirus particles that are capable of infecting a variety of human cells. The resulting defective virus is incapable of forming RCA, and it transmits up to 34 kb of foreign genes in the baggage area. Unlike conventional Ad5 vectors that require separate constructs for E. coli propagation of

insert genes, and recombination in vivo, the present vectors are relatively easy to make and provide a precise, safe alternative to first generation and second generation adenovirus vectors.

Exemplary methods for producing self-assembling vectors and genes are provided below. Further, the Examples provide methods for producing libraries of nucleic acid sequences using the methods of this invention. A number of nucleic acid sequences identified using the methods of this invention are described. The examples provided below are exemplary and not limiting. All references and publications provided herein are incorporated by reference into this disclosure.

10

**Example 1**  
**Three-Piece Gene Self-Assembly with 100% efficiency**

Using 6 primers (SEQ ID NOS:24 and 63-67), three PCR fragments were amplified from templates VLMG (SEQ ID NO:22) and VLBPGN (SEQ ID NO: 1). PCR reactions were carried out using the hot start technique, according to the manufacturer's instructions (Perkin Elmer) using *Pfu* DNA polymerase (Stratagene). To amplify specific portions of the above templates, each primer contained a class IIS enzyme site capable of digesting a unique overhanging end that was complementary to only one other terminus in the subsequent ligation. The class IIS enzymes used were *Bpm*1 and *Eco* 57I (the latter was used to copy a fragment that contained an internal *Bpm*1 site). The reactions were carried out as follows: 1) the lower reaction was assembled according to the protocol for PCR Gems (Perkin Elmer); 2) the lower reaction was heated to 80°C, 5 min, then cooled to 4°C for 5 min; 3) the upper reaction was prepared according to PCR Gems protocol and was added to the lower reaction (separated by cooled wax). The primer concentration was 0.3 μM (final). The dNTP concentration was 200μM (final). 5 Units of *Pfu* polymerase was used. All fragments were amplified using the following conditions: 96°C, 45 sec; (then followed by 30 cycles of the following) 96°C 45 sec, 52°C 45 sec, 72°C, 6 min; then followed by a single incubation at 72°C for 10 min; then hold at 4°C. All fragments were successfully amplified. The PCR fragments were purified using the Qiaquick PCR purification protocol (Qiagen). The fragments were digested with an excess of the appropriate restriction enzyme (*Bpm*1 or *Eco*57I). The digested fragments were run on a 1% agarose gel and were excised using minimal irradiation from a hand-held 365 nm ultraviolet light. The fragments were purified

using the Qiagen Qiaquick Gel Purification Protocol. The fragments were ligated at an equimolar ratio at a concentration of >20 $\mu$ g/ml with T4 DNA ligase (Boehringer Mannheim) overnight at 4°C. Competent *E. coli* SCS110 cells (Stratagene) were transformed with the ligated DNA. Eight colonies were characterized by restriction enzyme analysis, and all eight 5 contained the correct order and orientation of the three fragments. The experiment was repeated independently by another investigator, and the same result was obtained (8/8=100%). Thus, the procedure resulted in a high percentage of correctly assembled vectors.

This three-piece vector was VL $\Delta$ BP. The deletion extended from the distal 10 enhancer region to the TATA box near the start of transcription. The deletion region was a pair of *Bpm*1 sites that permitted U3 sequences to be cloned into the insert.

One validated *E. coli* clone of VL $\Delta$ BP was transfected into retroviral helper 15 cells. After 48 h, the vector was transduced into amphotropic helper cells. After selection for two weeks with the drug G418, drug resistant colonies were grown up in a mass culture and the vector was transduced from the amphotropic helper cells into a human HT1080 cell line 20 (ATCC, Rockville, MD). Surprisingly, even with a large deletion in the LTR promoter, the basal TATA box-containing VL $\Delta$ BP was transmitted as a retrovector and was permanently inserted into the human cell line, thus establishing the validity of the self-assembly technique for the construction of functional eukaryotic vectors.

20

### **Example 2** **Production of a Six Piece Self-Assembling Expression Vector**

Due to the high efficiency of the gene self assembly process for the three piece 25 assembly, a complex vector containing six fragments was constructed. The results here were extended to determine whether such a self-assembled vector would also have biological activity in human cells without being cloned and grown in a prokaryotic cell.

Six fragments were individually constructed by PCR using three different templates and twelve primers (as illustrated in Fig.8). The primers used three different class 30 IIS enzymes. The enzymes were chosen so as to give 2 base pair, 3'-overhanging ends. Three enzymes were used in order to avoid the use of enzymes that had additional sites internal to the fragments being amplified. Thus, *Bpm*1 was used unless there was an internal *Bpm*1 site. If such a site existed, *Eco*57I was used. If there was also an internal *Eco*57I site, then *Bsr*D1

was used. However, it is alternatively possible to use an enzyme such as *Eam*11041, where the *Eam*11041 sites in the primers are unmethylated (therefore susceptible to digestion by the enzyme), and wherein the <sup>m5</sup>dCTP analog of dCTP is used in the PCR reaction, methylating all internal sites (and protecting them from digestion by *Eam*11041), as suggested by Padgett and Sorge, 1996, *supra*, and incorporated herein by reference.

Using 12 primers, 6 fragments were amplified from 3 templates: pBK-CMV (SEQ ID NO:26) , pVLMB (SEQ ID NO:23) and pVLOVhGH-900 (SEQ ID NO:21). Fragment 1 was amplified from pBK-CMV using primers 1 and 2 (SEQ ID NOS:31 and 32). Fragment 2 was amplified from pVLMB using primers 3 and 4 (SEQ ID NOS:33 and 34). Fragment 3 was amplified from pVLOVhGH-900 using primers 5 and 6 (SEQ ID NOS:35 and 36). Fragment 4 was amplified from pVLMB using primers 7 and 8 (SEQ ID NOS:37 and 38). Fragment 5 was amplified from pVLMB using primers 9 and 10 (SEQ ID NOS:39 and 40). Fragment 6 was amplified from pVLMB using primers 11 and 12 (SEQ ID NOS:41 and 42). PCR reactions were carried out using the hot start technique, according to the manufacturer's instructions (Perkin Elmer Ampliwx PCR GEMS 100). The lower reaction was heated to 80 ° C for 5 min, then cooled to 20 °C for 5 min. The upper reaction was prepared according to PCR gems protocol and was added to the lower reaction (separated by cooled wax). The primer concentration was 0.3 micromolar (final). The dNTP concentration was 200 µM (final). 5 U of *Pfu* polymerase (Stratagene) was used per reaction. 100 ng of template was used for each reaction 14 rounds of PCR amplification were used to reduce mutagenesis of the templates. The PCR cycling protocol was 96 °C 45 sec; then two cycles of (96°C 45 sec, 52°C 45 sec, 72°C 6 min); then 12 cycles of (96°C 45 sec, 58°C 45 sec, 72°C 6 min) followed by a 72° C soak for 10 min, then to 4°C hold.

The six PCR fragments were designed to self-assemble into a retro-vector after digestion with the correct class IIS restriction enzyme (Fig. 8). After transfection into retroviral helper cells, the vector DNA is transcribed as RNA by means of the cytomegalovirus immediate early promoter (fragment 1). This promoter replaces the retroviral or VL30 LTR in this vector. The RNA transcript region begins with the R and U5 regions of the Moloney murine leukemia virus (MoMLV) LTR, the viral packaging signals ( $\Psi$ ) region of MoMLV, the packaging enhancer ( $\Psi^+$ ) region of mouse VL30 and the IRES region of EMCV fragment 2. Fragment 3 consists of the human growth hormone (hGH) cDNA sequence. Fragment 4 consists of the SV40 virus early region promoter driving

expression of the neomycin phosphotransferase (neo) gene. Fragment five consists of the (+)-strand primer binding site of the MoMLV LTR, the U3 region of the MoMLV LTR, the repeat (or R) region, and a portion of the U5 region. Fragment 6 consists of the PBR322 plasmid origin of replication.

5

Fragment 1: CMV early region promoter

Template: pBK-CMV plasmid DNA (Stratagene, LaJolla, CA) *Bpm1* (SEQ ID NO:26)

PCR primer 1 (SEQ ID NO:31)

10 GACTAACCTGATTCC~~ACTGGAGCCGTATTACCGCCATGCATTAGTTATTAATAG~~  
PCR primer 2 (SEQ ID NO:32)

GACTAACCTGATTCC~~ACTGGAGTAATTGCGGCTAGCGGATCTGACG~~

Fragment 2: R-U5-Psi-Psi(+)-IRES *Bpm1*

15 Template: pVLMB plasmid DNA (SEQ ID NO:23)

PCR primer 3: SEQ ID NO:33

GACTAACCTGATTCC~~ACTGGAGACACTTGACCTCTACCGCGCCAGTCCTCCGAT~~  
TGACTGAGTCG

PCR primer 4: SEQ ID NO:34

20 GACTAACCTGATTCC~~ACTGGAGGGATCCGCGCCCATGATTATTATCG~~

Fragment 3: human growth hormone (hGH) *Bsr* DI

Template: pVLCNOVhGH plasmid DNA (SEQ ID NO:21)

PCR primer 5: SEQ ID NO:35

GACTAACCTGATTCC~~CAGCAATGTCGGTTAGCTTGTTCAGTGTGTC~~

25 PCR primer 6: SEQ ID NO:36

GACTAACCTGATTCC~~CAGCAATGTTAGGACAAGGCTGGTGGGACTGG~~

Fragment 4: SV40 early promoter-neomycin phosphotransferase

Template: VLMB plasmid (SEQ ID NO:23)

30 PCR primer 7: SEQ ID NO:37

GACTAACCTGATTCC~~ACTGGAGGGTCGACCCTGTGGAATGTGTGTCAG~~

PCR primer 8: SEQ ID NO:38

GACTAACCTTGATTCCACT**GGAGAATCTCGTATGGCAGGTTGGCGT**

Fragment 5: MLV(+)PBS-U3-R-U5

5 Template: VLMB plasmid (SEQ ID NO:23)

PCR primer 9: SEQ ID NO:39

GACTAACCTTGATTCCACT**GAAGAGATTTATTTAGTCTCCAGAAAAAGGGGGG**

PCR primer 10: SEQ ID NO:40

GACTAACCTTGATTCCACT**GAAGCCCCAAATGAAAGACCCCCGCTGACG**

10

Fragment 6: PBR322 origin of replication

Template: VLMB plasmid (SEQ ID NO:23)

PCR primer 11: SEQ ID NO:41

GACTAACCTTGATTCCACT**GGAGCCGGACGGAATCGTAATCTGCTGC**

15

PCR primer 12: SEQ ID NO:42

GACTAACCTTGATTCCACT**GGAGTTCTCGAGGCGCGCATCTCGGCG**

20 *Procedure:* The twelve primers were prepared by the following procedure: 1) oligonucleotides were synthesized with trityls off. After deprotection and lyophilization, the samples were resuspended in 5 microliters deionized formamide and loaded onto a polyacrylamide gel (12% polyacrylamide, 250V). The samples were excised under short wave UV irradiation and eluted overnight in 600 microliters of sample elution buffer (0.5 M ammonium acetate, 10 mM Mg acetate, 1 mM EDTA, 0.1% SDS). The contents were loaded onto a BioRad Chromatography column (Cat. # 732-6008) and centrifuged into an Eppendorf tube at low speed (2000 RPM, 5 min). After washing the column with 500 microliters TE buffer (10 mM Tris, 1 mM EDTA), pH 8.0 and recentrifugation (2000 RPM, 5 min), the pooled eluate was ethanol precipitated, washed with 100% ethanol, resuspended in TE buffer and quantitated by spectrophotometry of a small sample, which was then discarded.

25

30 Fragments were cleaned using the Qiaquick PCR cleanup procedure. The fragments were digested with their respective class IIS restriction enzyme. The digested fragments were run on 1% agarose gels, and the fragments were excised and cleaned using the Qiaquick gel cleanup procedure. Fragments were combined in an equimolar mixture and

ligated overnight at 4° C with T4 ligase and ATP. An analytical gel was run with the ligated DNA, as well as with controls including unligated fragments and ligated fragments with a single fragment missing. As opposed to the controls, the complete ligation included bands equivalent to the full-length supercoiled monomer (referred to as GENSA 981, SEQ ID NO:29), as well as bands possibly representing multimers (up to six bands were observed).

In order to assess the efficiency of the method, eleven nanograms of DNA were transfected into SCS1 supercompetent cells. Thirteen kanamycin resistant colonies were harvested, and plasmid DNA preps indicated 10 out of thirteen that appeared to be the correct length. All ten gave the expected bands when digested with *Pst*1, *Sna*B1, and *Bam* HI. 1.35 µg of the ligated DNA was purified by phenol-chloroform-isoamyl alcohol extraction, followed by two extractions with chloroform-isoamyl alcohol, and was precipitated in ethanol. The DNA was washed in 70% ethanol and re-suspended in 50 µl of sterile phosphate buffered saline (for transfection). The DNA was transfected (using the Qiagen Superfect protocol) into HTAm1 (amphotropic human helper cells). 24 h after transfection, the target cells were washed and fresh culture media was added. 48 h after transfection, the supernatant from the vector producer cells was filtered (0.45 µm, Nalgene) and transferred to PG13 helper cells (ATCC) and HT1080 human fibrosarcoma cells. This procedure was repeated after 72 h. 48 h after transduction, recipient cells were started on G418 drug selection (500 µg/ml). The appearance of G418 drug-resistant colonies on transduced PG13 and HT 1080 cells after 6 days of selection indicated successful transmission via retrovirus particles. The transfect HTAm cells were also selected with G418. After six days of drug treatment, 45 colonies of resistant cells were counted. Thus, the six fragment gene assembly was effectively transmitted and expressed as either a DNA (transfection) vector or a retro-vector.

25

### **Example 3**

#### **Design and Construction of Single LTR Vectors**

*Background:* In order to manipulate the interior of the VL30 LTR sequences using a promoter rescue technique, single LTR vectors were constructed. The mouse VL30 element NVL-3 was used as the starting material as it is constitutively and abundantly expressed in most mouse tissues. Single LTR vectors are circular and behave as if they contained two LTRs. Thus, in these vectors RNA transcription begins at the start of the R region (see Fig.

3B), and continues through the polyadenylation site after completing the second round of transcription of the R sequences (Fig. 3A). In previous studies, these vectors were expressed transiently in vector producer cells and the DNA did not integrate into cell DNA as a standard two LTR vector. Therefore, the vectors were usually passed to a second complementation 5 helper cell line via retroviral transduction of the vector RNA transcribed in the first helper cell. This process resulted in the vector regenerating a correct (two LTR) structure upon integration into the recipient cell DNA.

*Experimental method:* The plasmid pNVL-3 (SEQ ID NO:25, kindly provided by Dr. J. Nortonm Manchester, UK), containing a complete copy of the NVL-3 (mouse VL30) genome (Adams *et al.*, 1989), was digested with *Xho*1 (which cuts in the LTRs), releasing the 4.27 kb VL30 genome with one copy of the LTR. This fragment was circularized using T4 DNA ligase and ATP. The circular DNA was linearized by digestion with *Sna*BI, 187 bp from the 3'-LTR. A 2.3 kb fragment containing the SV40 virus early region promoter and the 15 aminoglycoside phosphotransferase (*neo*) gene, together with the PBR322 plasmid origin of replication, was excised from the BAG retrovirus vector (Price *et al.*, *Proc. Natl. Acad. Sci.* 84:156-160, 1987, kindly provided by C. Cepko, Cambridge, MA). BAG is also obtainable in a retrovirus helper cell line from American Type Culture Collection (ATCC), Rockville, MD by digestion with *Xho*1 and *Bam*HI. This fragment was blunted with T4 DNA 20 polymerase and dephosphorylated with calf intestinal alkaline phosphatase (CIP). The fragment was then ligated to the linearized *Sna*BI fragment of NVL-3. The resulting plasmid (containing a circularly permuted NVL-3 genome with the SV-*neo-ori* region) was designated VLSNO2 (SEQ ID NO:30).

In order to facilitate the switching of LTR sequences by means of the class IIS 25 enzyme *Bpm*1, VLSNO2 was digested with *Bpm*1 (six sites). The region containing four *Bpm*1 sites was removed and replaced with a 19 bp linker (SEQ ID NOS: 1 and 52, see below), 921 bp beyond the LTR. The linker contained *Sna* BI, *Cla*1 and *Bam* HI cloning sites.

Linker (top strand): 5'-TACGTATCGATGGATCCGA-3' (SEQ ID NO:51)  
30 Linker (bottom strand): 5'-GGATCCATCGATACGTAAG-3' (SEQ ID NO:52)

The remaining two of the *Bpm1* sites had complementary ends, which permitted their ligation and resulted in eradication of all *Bpm1* sites within the resulting vector VLSNO3 (SEQ ID NO:20).

In order to facilitate reporter/therapeutic gene function, a 3.7 kb fragment containing the internal ribosome entry site (IRES) from encephalocarditis virus, together with the β-galactosidase reporter gene, was excised from the plasmid pVLSAIBAG (kindly provided by Mr. James Grunkemeyer, Omaha, NE) by means of a partial digestion of the plasmid with *Bam* HI. This region was inserted into the *Bam* HI site of VLSNO3, resulting in the vector VLSNOSIB (SEQ ID NO:14).

A second reporter construct, pVLSNOG (5774 bp, SEQ ID NO:19) contained the green fluorescent protein (GFP, Clontech, Palo Alto, CA) gene was constructed by inserting a *Bgl*2-*Bcl*1 fragment (800 bp) from plasmid pGFP-N1. This sequence, containing the GFP gene, was treated with mung bean exonuclease and inserted into the unique *Sna* B1 site of pVLSNO3.

In order to enhance GFP fluorescence from the reporter plasmid pVLSNOG, the serine-65 codon in the GFP gene was mutated into threonine by a site-directed mutagenesis procedure with the Transformer™ Site-Directed Mutagenesis kit from Clontech. A *Bpm1* site in the GFP gene (threonine-9) was mutated at the same time without changing the amino acid (ACT to ACA). The resulting plasmid was pVLSNOGM (SEQ ID NO:18).

An *Nco*1-*Xho*1 fragment (585 bp) from plasmid pG1IL2EN (kindly provided by Dr. Steven Rosenberg, Bethesda, MD), containing the internal ribosome entry site (IRES) from encephalomyocarditis virus (EMCV) was inserted into the *Apal* site upstream of the GFP gene in pVLSNOGM, resulting in pVLSNOGMI (SEQ ID NO:17). Both insert and plasmid fragments were blunted with mung bean exonuclease. One variant version of pVLSNOGMI with an IRES tandem dimer was also constructed and designated pVLSNOGMI2 (SEQ ID NO:16).

Oligonucleotides (SEQ ID NO:53 and 54) containing a splice acceptor (SA) of AKV virus (in bold) was inserted into pVLSNOGMI at the unique *Sac* 2 site just before the IRES, resulting in pVLSNOGMIS (SEQ ID NO:15).

Oligo: (SEQ ID NO:53)

5' -GGCCG**CTAACTAATAGCCCATTCTCCAAGGTACGTAGC**- 3'

3' - CGCCGGCGATTGATTATCGGGTAAGAGGGTTCCATGCAT - 5'

(SEQ ID NO:54, bottom Oligo)

### Recovery of LTR promoter sequences from mouse CD4+ T-helper cells

In order to facilitate the recovery of VL30 promoter sequences expressed in mouse T-helper cells, a mouse CD4+ T-helper cell cDNA library (Stratagene, San Diego, CA, Catalog # 937311) was screened by plaque hybridization. Approximately  $2 \times 10^4$  bacteriophage  $\lambda$ -ZAP clones were plated on a lawn of *E. coli* cells according to the manufacturer's instructions. Two nylon filters were sequentially layered onto the lawn of *E. coli* cells and bacteriophage. The filters were hybridized to a  $^{32}\text{P}$ -labelled (Prime-It RmT Random Primer Labeling Kit, Stratagene), 4.2 kb internal *Xho*1 fragment of NVL-3 (containing the NVL-3 genome). 55 plaques (or approximately 0.3% of the total phage) reacted positively on both filters. 18 VL30 cDNA sequences were cloned from the plate, which was used to identify U3 promoters that are actively expressed in the RNA of mouse T-cells. Five of the 18 clones contained intact U3 sequences, representing four of one molecular species, named TH1 (SEQ ID NO: 2) and one of another species, named TH2 (SEQ ID NO: 3) also provided in Fig. 5. TH1 contained approximately 120 bp more DNA than did TH2. Because TH1 was more abundant (4 out of 5 clones), the additional sequences in the enhancer region were implicated to be a possible reason for the stronger expression in mouse T cells. Examination of the known and putative transcription factor binding sites in the VL30 LTR (Hodgson, 1996, chapter 4, Fig. 4.2 *supra*) revealed several interesting features of TH1 and TH2. First, the extra sequences of TH1 that were missing in TH2 included an extra copy of the enhancer repeat region as well as a potential retinoid (RAR/RXR) binding site. Several transcription factor binding sites in the enhancer repeat region that differed between the two elements included: a cyclic 3'-5'AMP response element (VLCRE, a potential CREB/jun binding site), a serum response element (SRE), and a potential NF1/IL6 binding site (although there were additional sites for these factors in other enhancer repeats). These factors could possibly explain why VLTH1 appeared to be expressed at higher levels, both in the source cells and into transduced cells. Together, the VL30 sequences represented 0.3% of the mRNA expressed in the T cells, and TH1 appeared to be most abundant VL30.

**Sequencing Primers:**

(SK, SEQ ID NO:49) 5'-CGCTCTAGAACTAGTGGATC (20 mers, Tm 60°C).

(T7, SEQ ID NO:50) 5'-GTAATACGACTCACTATAGGG (21 mers, Tm 60°C).

5       **Seamless Rescue of T cell promoters using class IIS restriction enzymes**

Two sets of primers containing offset *Bpm*1 restriction sites were designed and synthesized. One set was for amplification of the plasmid sequences, and another was for the amplification of the inserts.

10       **Insert Primers: (*Bpm*1 site bold)**

ITA (43 mer, Tm: 67.2 °C, SEQ ID NO:45)

CGATCCACT**GGAG**CTCGGAGCCCACCCCCCTCCCATCTAGAGGT

## 15       ITB (43 mers, Tm: 66.3 °C, SEQ ID NO:46)

CGTCCTCCT**GGAG**AGCACAGGGTAGAGGAGTCTCGACGGTCAG**Vector primers: (*Bpm*1 site bold)**

VLA (43 mers, Tm: 68.2 °C, SEQ ID NO:47)

CGCAACC**CTGGAG**ACCTCTAGATGGGAGGGGTGGGCTCCGAG

## 20       VLB (43 mers, Tm: 66.3 °C, SEQ ID NO:48)

GCAGGAC**CTGGAG**CTGACCGTCGAGACTCCTCTACCCCTGTGCT

To amplify vector sequences more efficiently, vector templates were shortened by deleting marker genes from vectors. pVLSNOSIB (SEQ ID NO:14) was cut with *Kpn* 1 and a 4201 bp fragment containing β-gal gene was removed. The remaining vector has 3923 bp.

The U3-promoter inserts (357 bp for TH1 and 240 bp for TH2) were PCR-amplified from TH1 and TH2 promoters with primers ITA and ITB. The vector cassettes (~4.2 kb for pVLSNOSIB and ~3.7 kb for pVLSNOGMIS) were PCR-amplified from the shortened vector templates using primers VLA and VLB, (*supra*). The PCR-amplification was done with high-fidelity *Pfu* DNA polymerase from Stratagene (La Jolla, CA). The amplified products were gel-purified (1% agarose gel). The inserts were then cut with *Bpm* 1 to produce complementary ends. The vector cassette products were phosphorylated with

PNK, then circularized with T4 ligase, and transformed into SCS 110 cells. Recovered plasmids were then digested with *Bpm* 1 and treated with CIP to produce complementary ends. *Bpm* 1 treated inserts and vector cassettes were ligated, and T-cell tissue-specific VL 30 vectors VLTH1 and VLTH2 were produced. The marker  $\beta$ -gal gene and GFP gene were put back into those vectors at the original unique sites *Kpn* 1 and *Sal* 1 respectively.

#### **Transmission and expression of single LTR vectors and T cell U3 sequences**

Vector DNA constructs were transfected into GP+E86 retroviral helper cells (Markowitz et al, 1988, *supra*) using the Lipofectamine protocol (Life Technologies, Gaithersburg, MD). The culture media from these cells (supernatant), containing defective transducing particles (72 h post-transfection), was transmitted to PA317 (Miller, US Patent, cited *supra*) amphotropic helper cells, using Lipofectamine to enhance transduction efficiency (Hodgson et al., 1996. Synthetic Retrotransposon Vectors and Gene Targeting pp. 3-14, in : Felgner et al., eds. *Artificial Self-Assembling Systems for Gene Delivery*. American Chemical Soc. Books, Washington, D.C.). A similar procedure was used to transmit VLTH1 and VLTH2 to the PG13 helper cell line (Miller et al., 1991. *J. Virol.* 65:2220-2224). 24 h post-transfection, the recipient cells were selected with the drug G418 (500 $\mu$ g/ml, 2 weeks) to enrich for stably transduced cell populations.

All of the single LTR vectors, including VLTH1 and VLTH2 were transmitted by this method, indicating that single LTR vectors can be used for promoter switching and yet revert to dual LTR vectors after a single passage. Vectors VLSNO2, VLSNO3, and VLSNOSIB were then titered on NIH 3T3 cells (using the PA317 vector producer cell lines). VLTH1 and VLTH2 vectors were titered on human HT1080 cells (PG13 cell lines). Surprisingly, all of the single LTR vectors were transmitted effectively. However the titers of stably transduced TH1 and TH2 cell lines were  $5.5 \times 10^2$ - $1.1 \times 10^3$  TU/ml, compared to  $0.4$ - $3.0 \times 10^4$  TU/ml for the VLSNO2, VLSNO3 and VLSNOSIB cell lines. Thus, switching from the NVL-3 transcriptional promoter (originally isolated from NIH 3T3 fibroblast cells) to VL30 promoters derived from T helper cells, appeared to have a negative effect on RNA expression in fibroblast cells, as determined by the transmissibility of the RNA.

In order to study the usefulness of rescued promoters as DNA transfection vectors (as opposed to retro-vectors), VLSNOSIB, VLTH1 and VLTH2 were also transfected into a number of cell lines (using Lipofectamine), including NIH 3T3, PA317, GP+E86,

PG13, HT1080, SW480 and HeLa (available from ATCC). RNA expression in these cell lines is shown in Table 4, wherein gene expression from the LTR promoter (as determined by  $\beta$ -gal staining) is normalized to VLSNOSIB (100).

<b>Cell line:</b>	<b>NIH 3T3</b>	<b>PA317</b>	<b>GP+E86</b>	<b>PG13</b>	<b>HT1080</b>	<b>SW480</b>	<b>HeLa</b>
<b>Vector:</b>							
<b>VLSNOSIB</b>	100	100	100	100	100	100	100
<b>VLTH1</b>	39.3	18.7	0.1	21	25.5	156	156
<b>VLTH2</b>	28.6	7.1	5.5	11.5	46.8	82	156

5           Table 4. Transient expression of a  $\beta$ -gal marker gene by three VL30 promoters: NVL-3 (VLSNOSIB), VLTH1 and VLTH2. Cells were transfected using the Lipofectamine procedure. Total blue cells were counted from each well in 6-well plates, and the number of blue cells from VLSNOSIB was normalized to 100%.

10

The expression of both the VLTH1 and VLTH2 promoters was significantly reduced compared to VLSNOSIB in cell lines of fibroblastic origin, whereas in SW480 colorectal cancer cells and HeLa cells, it was comparable to or better than VLSNOSIB (the NVL-3 promoter). However, VLSNOSIB was expressed poorly in the non-fibroblastic cell lines, so a direct comparison was difficult to interpret. Unfortunately, the human T cell lines (Jurkat and MOLT4 [obtained from ATCC]) were not transfected by Lipofectamine, and they were poorly transduced by VLTH1 and VLTH2 retro-vectors. In the Jurkat and MOLT4 cells transduced with VLTH1 and VLTH2, only a small percentage (1-10%) of cells that were 15 stably transduced by the vectors stained positively for  $\beta$ -gal expression. However, the marker gene (neo) continued to be expressed from an internal promoter, as evidenced by drug 20 selection.

Taken together, the results demonstrated: 1) the ability of the promoter rescue technique to seamlessly capture functional transcriptional promoters from specialized cells; 2) 25 the ability of single LTR vectors to introduce the rescued promoters into standard transducing vectors; 3) the ability of the rescued promoters to be expressed at differing levels in several different cell types, including T cells; and 4) screening and selection established the efficacy, or lack thereof, of individual promoter sequences.

Although the general method of promoter rescue was demonstrated by the 30 foregoing experiments, the titers obtained from the sLTR VL30 vectors may not be useful where selection systems are not available.

Additional experimentation led to the development of a chimeric packaging signal, combining the essential packaging signal from Moloney murine leukemia virus ( $\Psi$ ), and the enhanced packaging signal ( $\Psi+$ ) from a mouse VL30 element. A vector embodiment of this packaging system is VLMB (SEQ ID NO:23). One advantage of the chimeric 5 packaging system was the elimination of retroviral *gag* gene sequences that were present in previous high-titer MLV-based vectors (viral *gag* sequences contribute to the generation of replication competent retrovirus outbreaks). The titers of VLMB-based vectors ranged from approximately  $1 \times 10^5$  to  $4 \times 10^6$  TU/ml.

10           **Construction of a cloning vector for promoter rescue**

Using pVLSNOGMIS as a template, and primers (SEQ ID NOS:28 and 68), a 6.4 kb plasmid fragment was PCR amplified (Using Hot Start AmpliTaq PCR Gems 100, Perkin Elmer). 30 cycles of PCR were performed by following the manufacturer's instructions, with the following input conditions: lower reaction, 80° C, 5 min., then add 15 upper reaction and template, 96° C, 1 min. Each reaction vial contained 50 ng template, 0.5  $\mu$ M each primer, 200  $\mu$ M dNTPs and 5U (2 $\mu$ l) *Pfu* polymerase (Stratagene, LaJolla, CA). 30 repeating cycles of: 96° C, 45° sec; 50° C, 45 sec; 75 C, 1 min. A final incubation of 75° C, 10 min, then hold at 4° C. After amplification, the reactions were purified using Qiaquick 20 PCR Purification Kits (Qiagen). The PCR products were digested with *Pac*1, heat inactivated (65° C, 20 min) and ligated together using T4 DNA ligase (overnight at 4° C in a 5  $\mu$ l vol). The ligated DNA was transfected into SCS110 *E. coli* cells (Stratagene) with kanamycin (50  $\mu$ g/ml) antibiotic added to the agar plates. The cells were *dcm*<sup>-</sup>, *dam*<sup>-</sup> (to prevent methylation of *Bpm*1 sites). The resulting plasmid, pVLBPGN (SEQ ID NO:1, Figs 2 &3) has a deletion 25 in the U3 region of the LTR. A linker containing a central *Pac*1 site flanked by two outwardly-digesting *Bpm*1 sites occupies the site of the deleted U3 sequences. The *Bpm*1 sites enable the plasmid to be digested with *Bpm*1, resulting in two 2 bp 3'-overhanging ends that are complementary to the U3-derived RT-PCR inserts described below. The digested plasmid was purified free from the intervening linker sequences from an agarose gel after digestion with *Bpm*1, using the Qiaquick gel purification kit (Qiagen).

**Procedure for amplification of liver U3 promoter region**

Purified mouse liver total tissue RNA was purchased from Ambion, Inc., (Austin, TX). Total liver RNA was treated with RQ1 Rnase-free (Promega, Madison, WI). Using Perkin Elmer Gene Amp thermostable rTth reverse transcriptase RNA PCR kit (P/N 5 N808-0069), the following conditions for RT-PCR were used: RT-PCR A 70° (hot start); RT-PCR B, 95°C, 60 sec, then 35 cycles (95°m 10 sec, 58°C, 15 sec) then a final 58°C incubation for 7 min, then 4°C and hold. Additional conditions were: primer concentration 0.15 micromolar, template 100 ng/reaction, dNTPs 200 micromolar (final) and MgCL<sub>2</sub> 3.5 mM(final). The primers for insert amplification were SEQ ID NOS:28 and 68)

10 The amplified U3 sequences were purified using Qiaquick. The pVLBPGN plasmid was digested with *Bpm*1, isolated from a 1% agarose gel and purified using the Qiaquick method. The purified U3 sequences were ligated at 1:2, 1:4 and 1:6 molar ratios of VLBPGN plasmid:insert using T4 DNA ligase and a 5 microliter reaction volume overnight at 4°C (100 ng plasmid: 16 ng insert = 1:1 molar ratio). 1 microliter of each ligation reaction 15 was transformed into *E. coli* SCS 110 competent cells (Stratagene). 26 colonies were recovered in total. Out of 23 clones grown overnight in the presence of kanamycin, 20 had sequences that appeared to be mouse VL30 sequences, representing 10 different VL30 species (Fig. 6, SEQ ID NOS: 4-13). One of these (Hep 10, SEQ ID NO: 13) was transiently transfected into Hep G2 liver hepatocellular carcinoma cells. 48 h after transfection, intense 20 GFP fluorescence was observed, indicating strong expression of the Hep 10 U3 promoter region.

**Example 4  
Creating a combinatorial library of mouse VL30 U3 sub-regions.**

25 Using Fig. 7 and Hodgson, 1996, supra, Fig. 4.2 as a guide, the following three sub-regions of the VL30 U3 region were empirically established: Distal (1); medial (2); and proximal (3). Peaks of similarity were used to guide the following choice of primers: (+) primer binding site-5'-LTR boundary; ~80 bp (defines sub-region 1); ~80-210 bp (sub-region 2); ~210-430 (sub-region 3). The following primers were selected to amplify the vector VLBPGN or a similar VL30, NVL-3 LTR-containing vector:

P1 (going left from the 5'-end of the LTR to amplify the plasmid)

(SEQ ID NO:55)

GACTAACCTTGATTCCACTGGAGTTT(CT)(CT)ATTCTCATTCCCCACTTC  
TTCTT

P2 (going right from the 3'-end of the promoter region to amplify the plasmid)

5

(SEQ ID NO:56)

GACTAACCTTGATTCCACTGGAGAATCTGGACCAATTCTATATAAGCCTG  
TGAAAAATT

The six primers selected to amplify the inserts are as follows:

10

Fragment 1, primer 1 (going right from the LTR terminus into U3) (SEQ ID NO:57)

GACTAACCTTGATTCCACTGGAGAAGAAGAAGAAGTGGGAATGAAGAA

Fragment 1, primer 2 (going left from the end of fragment 1) (SEQ ID NO:58)

GACTAACCTTGATTCCACTGGAGATCTCTAGATGGGAGGG(GT)(CT)GGG  
CTC

15

Fragment 2, primer 1 (going right from the left end of fragment 2) (SEQ ID NO:59)

GACTAACCTTGATTCCACTGGAGCTGGAGGCCACCCCCTCCCATCT

Fragment 2, primer 2 (going left from the right end of fragment 2) (SEQ ID NO:60)

GACTAACCTTGATTCCACTGGAGGGAGGCCCTATCTCAAAAATGTT

Fragment 3, primer 1 (going right from the left end of fragment 3) (SEQ ID NO:61)

20

GACTAACCTTGATTCCACTGGAGTCTAAGAACATTTGAGATAAGGGCC  
T

Fragment 3, primer 2 (going left from the right end of fragment 3) (SEQ ID NO:62)

GACTAACCTTGATTCCACTGGAGTCACAGGCTTATATAG(TG)AAA

25

100 ng of genomic DNA from *Mus musculus* is used as a template (the mouse genome

bears 100-200 copies of VL30 elements). Standard PCR procedures for *Pfu* polymerase are

used. Fragments are amplified 35 rounds of PCR to obtain single-copy genomic DNA

amplification. Samples of Qiagen column purified DNA are examined on analytical agarose

gels to determine the approximate size. The remainder of each reaction is digested with the

30

appropriate enzyme and run on an acrylamide or agarose gel. The digested fragments are

purified by standard gel purification procedures and are ligated to the plasmid fragment at an

equimolar ratio of the four PCR fragments (three inserts and one plasmid). The ligation mix

is transformed into *E. coli* SCS1 and is grown on kanamycin. The number of colonies is used to establish the size of the combinatorial library, and the pooled colonies are grown in *E. coli* and the DNA is harvested *en masse*. A dozen or more colonies are characterized by DNA sequencing to determine the approximate fidelity of the reaction. A library of 1,000 or more, 5 but preferably 100,000 or more members is used for combinatorial screening procedures.

**Screening the combinatorial libraries for expression in specific cell types using a replication defective helper virus**

The U3 library DNA is transfected into the desired target cells in which expression is desired. Along with the library, approximately 25% of the total DNA should include retroviral helper sequences. The latter sequences can be a helper plasmid (such as pPAM3, Miller *et al.*, US Patent 4,861,719). The virus is amphotropic, permitting it to infect most human cells. The RNA from individual clones that are transcribed in the target cells will be packaged into retroviral virions made by the helper virus, and the virions can be harvested 10 as the cell free filtrate (0.45 mm) from the vector producer cells. This virus (containing the expressed sequences) can be transmitted to fresh target cells that do not contain helper virus. 15 48 hours after passage, the DNA form of the transcriptionally active clones will be integrated in the recipient cells, and these transcriptionally active loci will produce more RNA, and protein. After G418 drug selection to increase the proportion of cells expressing the vector 20 sequences, helper virus DNA is again transfected into the recipient cells, transforming them into vector producer cells. The virus from these cells should contain increased amounts of the RNA from clones that are transcriptionally active in those cells. Passage of the virus is continued for two or three rounds to permit recombination and mutation to take place, enhancing the effect of *in vitro* evolution of promoters. The actual degree of enhancement 25 attainable at each step is illustrated in Table 2 (*supra*). After several passages, the actual level of RNA expressed by several clones is determined by RNA blotting, or by the amount of a reporter gene expressed as protein (determined visually or by the appropriate assay). Because human cells do not naturally contain VL30 DNA or RNA, the sequences that remain 30 in the human cells are those with the most transcriptionally active promoters. These sequences can be amplified and re-cloned using the methods of the instant invention, or they can be rescued by virus packaging, reverse transcribed by the endogenous reverse

transcriptase reaction, and grown as plasmids (due to their plasmid origin of replication and the selectable kanamycin marker gene).

In addition to using a replication defective helper virus, such as the clone pPAM3, it is also possible to use a replication competent retrovirus, such as Moloney murine leukemia virus to passage the library. For use in human cells, however, the virus should have a tropism that is compatible with human cells (gibbon ape leukemia virus and amphotropic [4070A] murine retroviruses are acceptable).

In addition to being useful for generating active transcriptional promoters *de novo*, a small variation on the above procedures may enable the isolation of hormone responsive promoters. In it, the cells are treated with the hormone (which could be a steroid, a peptide hormone known to affect the cells, a drug, a drug agonist or antagonist, etc.) during passage. After isolation of surviving VL30 vector-containing cells, individual clones of drug resistant cells are tested for reporter gene expression with and without drug treatment to determine relative protein expression. Likewise, RNA expression can be determined by blot analysis or a similar method. A useful list of known VL30 responses to pharmacological agents is listed in Fig. 4.2 of Hodgson, 1996, *supra*, and can be used as a guide to help assess the potential agents known to have an effect on VL30 transcription.

Once the transcriptional promoters with the known specificity have been obtained, they can be used to obtain expression of genes from a variety of types of vectors. For example, in addition to retrovirus particles, the promoters can be incorporated into all other major groups of vectors: adenoviruses, herpes simplex virus vectors, DNA transfection vectors, etc. It will be apparent to persons of ordinary skill in the art that similar combinatorial libraries can also be used to screen for other characteristics than transcription activity in a particular cell. For example, combinatorial libraries of complementarity determining regions (CDRs) of antibodies or T cell receptors can be so screened using antibody screening methods, such as the phage display screening method (Pharmacia, Milwaukee, WI). Thus, the methods of this invention, particularly the combinatorial simplicity of this invention is a significant improvement over many *in vivo* recombination methods including those of (Stemmer, US Patent 5,605,793; 1997) that have described for the production of CDR combinatorial libraries.

**Example 5**  
**Gene Assembly Line**

5           From the above examples of 3 and 6 fragment gene self-assemblies, it is evident that assembly of genes by means of gene amplification, the use of offset restriction enzymes and incorporating unique, non-palindromic ends is a highly efficient process compared to conventional cloning methods. However, in addition to the considerations already discussed, it will be apparent to a person of ordinary skill in the art that the various  
10          procedures, protocols, methods and material of the instant invention become more difficult to use as the number of fragments increases. For example, if the efficiency of combining each fragment in an assemblage is 99%, then the overall efficiency of combining ten fragments will be 90%, the efficiency of combining 100 fragments will be 37%, etc. Therefore, a small drop in efficiency of any step or fragment, or a large increase in the complexity of the project,  
15          will be sufficient to reduce the overall efficiency. Fastidious procedures permit one to achieve success with more complex projects.

Foremost in its potential for inducing failure is human error in primer design where large numbers of fragments are used. Fortunately, the instant invention is suited to automation of most of the steps. This allows human input to be focused on design, analysis,  
20          and quality control. For the purposes of generating large vectors or chromosomes, it is desirable to provide an automated environment. One method to achieve this goal is a gene assembly line.

In a gene assembly line, multiple tasks are controlled by a machine or machines working together to increase speed and efficiency and to reduce human error. For  
25          example, computer aided design (CAD) and computer aided manufacturing (CAM) are incorporated and combined with the methods of this invention. The computers accept inputs in the form of template and primer sequences, together with preferences of regions to be copied and joined. The preferences include at least the sequences of the primer regions and information about the known restriction sites and maps of the sequences to be assembled, but  
30          ideally include the entire sequence. The preferences also include the number of sequences to be joined, the desired Tm for the primers, the list of potential restriction enzymes capable of offset digestion that are potential candidates for use in the assembly process, the desired end structures for each fragment terminus, a tag sequence (if any), whether circular or linear ends

are desired, and additional design considerations. The computer algorithm then searches the sequences to determine the candidate enzymes and specific primers that match the criteria of the input. Candidates for selection of unique non-palindromic overlaps are selected. The computer then posts selections or preferences for the type and order of end structures, the 5 primer binding sites, their potential for primer-dimer and intra-molecular interaction artifacts, and the potential conflicts with repeat sequences within the templates that could lead to incorrect polymerization. Based upon the selections made by the operator, the computer then determine the  $T_m$  for each primer, and makes adjustments (with suitable inputs from the investigator) to achieve a suitable  $T_m$  for the appropriate DNA synthesis or gene amplification 10 reaction. Ideally, the primers should have similar  $T_m$ 's so that all amplification reactions can be performed at once with one set of amplification instructions. In reality, it may be difficult to do this with complex projects. The output of this portion of the program, which can be in a generic format, such as a Microsoft Excel spreadsheet is then downloaded to a computerized oligonucleotide synthesizer, such as the Applied Biosystems 3928 nucleic acid synthesizer. 15 One advantage of using a computerized synthesizer is its robotic capability to de-protect and purify the oligonucleotides automatically. In addition this synthesizer can accept computerized input.

The quantity of individual oligos recovered is then determined spectrophotometrically. It is desirable to purify the oligonucleotides by high performance 20 liquid chromatography or by polyacrylamide gel. In a preferred embodiment, the oligonucleotides and templates are then assembled robotically using an automated nucleic acid handling system such as the Qiagen BioRobot 9600. The BioRobot is capable of accepting input from a computer and can combine the gene amplification reactions based upon the assignments of templates, primer and reagents provided in the input. The assembled 25 reactions are then amplified for example by PCR. In a preferred embodiment, the PCR heat block is incorporated into the robotic workspace and genes are assembled robotically but with minimal human intervention to change buffers, rearrange the platform, change programs, and the like. The resulting amplified products are also purified by the BioRobot or a similar robotic device. In a preferred embodiment, the robotic device uses Qiaquick cleanup 30 procedures, or a similar method and then assembles restriction endonuclease reactions to digest the purified gene amplification products. The gene amplification products are loaded onto a gel and electrophoresed. Human intervention may be necessary to analyze the

products and excise the correct fragments from the gel. At this point, the results are assessed and missing or incorrect sized fragments are resynthesized. The robotic device is preferably used to purify the gel fragments using Quiagen or similar cleanup procedures. After spectrophotometric quantitation of the purified fragments, the robotic device is preferably used to assemble the ligation. Ideally the fragments are combined in an equimolar ratio of 1:1. However it is not necessary to use equimolar ratios in order to achieve gene self-assembly. For automated gene assembly, it may be desirable not to use equimolar ratios of input fragments, particularly if it simplified the task of quantitation. After ligation, the assemblies can be purified and ethanol precipitated or they can be added to the appropriate host cells. Automation aids in maintaining the sterility of the reaction.

Several additional considerations can assist in the construction of long genes using gene assembly. First the number of fragments and the length of constructs are limiting factors. In addition to maintaining high standards of purify of both the oligonucleotide primers and gene amplification products, it is important to keep the error rate low during copying. Thus, one can optimally start with 100 ng of template use only five rounds of gene amplification and finish with nearly 2 micrograms of product. This is more desirable for reducing errors than using a large number of amplification steps. It is also desirable to use a special copying enzyme such as *Pfu* DNA polymerase that has a low intrinsic error rate. Further it is desirable to use *in vivo* selection (in eukaryotic cells or tissues) rather than *E. coli* cloning to reduce the incorporation of errors into the vectors. For example, a viral vector such as an adenoviral vector or the retro-vectors of the preceding examples are auto-selecting. A single correctly-assembled adenovirus vector molecule, for example, leads to a lytic infection (the viral products of which are cloned by limiting dilution on the appropriate eukaryotic cells), even though it may be combined in a ligation mix with a large excess of incorrectly assembled molecules that are non-functional. Thus, it is not necessary to have a high efficiency, although high efficiency has been demonstrated in this system, in order to achieve success in making, for example gene therapy vectors.

For long fragments (3-30 kb), it is desirable to use enzymes and procedures that are designed or facilitate replication of long fragments, one such example is the eLONGase system (Life Technologies). This system can copy up to 30 kb on a fragment with proofreading. Considerations for long PCR are reviewed in Beck, 1998. (The Scientist 6 January, 1998, pp. 16-18).

Internal restriction sites are a potential problem, particularly with large constructs and can be overcome in a number of ways. Use of alternate enzymes, methylation of internal restrictions sites (such as by using methylated DNA precursors during synthesis to leave the sites in primers unaffected, incorporation of the internal sites into the construct (if they are non-palindromic), or mutagenesis of internal sites, are exemplary ways to deal with some of these issues.

For very large constructs, it is desirable to use enzymes such as *SapI* (recognizing 7 nucleotides and leaving a 3 bp overhang). This enzyme digests every 16,384 bp on average. There are 64 nucleotide triplet combinations, meaning that up to 32 fragments can be ligated in a circle using *SapI*. *Fok1* and *Hga1* are other examples of class IIS enzymes that are useful for making large constructs. *Hga1* has 5 bp overhangs, permitting more than 500 *Hga1* fragments to be ligated. *Fok1* includes a Kozak ATG start codon. In a preferred embodiment, a *Fok1* site is inserted at the PuXXATG start site of a cDNA encoding region. The cDNA is inserted in frame, providing a site for inserting and switching coding sequences within a vector.

It will be readily understood by those skilled in the art that the foregoing description has been for purposes of illustration only and that a variety of embodiments can be envisioned without departing from the scope of the invention. Therefore, it is intended that the invention not be limited except by the claims.

## SEQUENCE LISTING

## (1) GENERAL INFORMATION:

5 (i) APPLICANT: NATURE TECHNOLOGY CORPORATION, ET AL.

(ii) TITLE OF INVENTION: SELF-ASSEMBLING GENES, VECTORS AND USES THEREOF

10 (iii) NUMBER OF SEQUENCES: 68

## (iv) CORRESPONDENCE ADDRESS:

- (A) ADDRESSEE: MUETING, RAASCH & GEBHARDT, P.A.  
(B) STREET: 119 NORTH FOURTH STREET, SUITE 203  
15 (C) CITY: MINNEAPOLIS  
(D) STATE: MINNESOTA  
(E) COUNTRY: USA  
(F) ZIP: 55401

## 20 (v) COMPUTER READABLE FORM:

- (A) MEDIUM TYPE: Floppy disk  
(B) COMPUTER: IBM PC compatible  
(C) OPERATING SYSTEM: PC-DOS/MS-DOS  
25 (D) SOFTWARE: PatentIn Release #1.0, Version #1.30

## (vi) CURRENT APPLICATION DATA:

- (A) APPLICATION NUMBER: Not Assigned  
(B) FILING DATE: 28-FEB-1998  
30 (C) CLASSIFICATION:

## (vii) PRIORITY APPLICATION DATA:

- (A) APPLICATION NUMBER: 60/070,910  
(B) FILING DATE: 28-FEB-1997  
35 (C) CLASSIFICATION:

## (viii) ATTORNEY/AGENT INFORMATION:

- (A) NAME: MCCORMACK, MYRA M.  
(B) REGISTRATION NUMBER: 36,602  
40 (C) REFERENCE/DOCKET NUMBER: 228.00010201

## (ix) TELECOMMUNICATION INFORMATION:

- (A) TELEPHONE: 612-305-1225  
45 (B) TELEFAX: 612-305-1228

## (2) INFORMATION FOR SEQ ID NO:1:

## (i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 6225 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
50 (D) TOPOLOGY: linear

55 (ii) MOLECULE TYPE: DNA (genomic)

60 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:1:

TGAAGAATAA AAAATTACTG GCCTCTTGTG AGAACATGAA CTTTCACCTC GGAGCCCACC 60

65 CCCTCCCATC TGGAAAACTC CAGTTATAAC TGGAGTTTTT CCTTTAAAG CTTGTGAAAA 120  
ATTTGAGTCG TCGTCGAGAC TCCTCTACCC TGTGCAAAGG TGTATGAGTT TCGACCCAG 180

	AGCTCTGTGT GCTTCTGTT GCTGCTTAT TTGACCCCCA GAGCTCTGGT CTGTGTGCTT	240
5	TCATGTCGCT GCTTTATTAA ATCTTACCTT CTACATTTA TGTATGGTCT CAGTGTCTC	300
	TTGGGTACGC GGCTGTCCCG GGACTTGAGT GTCTGAGTGA GGGTCTCCC TCGAGGGTCT	360
	TTCATTTGGT ACATGGGCCG GGAATTCGAG AATCTTCAT TTGGTGCATT GGCGGGAAAT	420
10	TCGAAAATCT TTCATTTGGT GCATTGGCCG GGAAACAGCG CGACCACCCA GAGGTCTTAG	480
	ACCCACTTAG AGGTAAGATT CTTTGGTCTG TTTTGGTCTG ATGTCGTGT TCTGATGTCT	540
15	GTGTTCTGTT TCTAAGTCTG GTGCGATCGC AGTTTCAGTT TTGCGGACGC TCAGTGAGAC	600
	CGCGCTCCGA GAGGGAGTGC GGGGTGGATA AGGATAGACG TGTCCAGGTG TCCACCGTCC	660
	GTTCGCCCTG GGAGACGTCC CAGGAGGAAC AGGGGAGGAT CAGGGACGCC TGTTGGACCC	720
20	CTTTGAAGGC CAAGAGACCA TTTGGGGTTG CGAGATCGTG GGTCGAGTC CCACCTCGTG	780
	CCCAGTTGCG AGATCGTGGG TTGAGTCCC ACCTCGTGT TTGTTGCGAG ATCGTGGGTT	840
25	CGAGTCCCAC CTCGCGTCTG GTCACGGAT CGTGGGTTCG AGTCCCACCT CGTGTGTTGT	900
	TGCGAGATCG TGTTGGTCTGAG TCCCACCTCG CGTCTGGTCA CGGGATCGTG GGTCGAGTC	960
	CCACCTCGTG CAGAGGGTCT CAATTGGCCG GCCTTAGAGA GGCCATCTGA TTCTTCTGGT	1020
30	TTCTCTTTT GTCTTAGTCT CGTGTCCGCT CTTGTTGTGA CTACTGTTT TCTAAAATG	1080
	GGACAATCTG TGTCCACTCC CCTTTCTCTG ACTCTGGTTC TGTCGTTGG TAATTTGTT	1140
35	TGTTTACGTT TGTTTTGTG AGTCGTCTAT GTTGTCTGTT ACTATCTTGT TTTTGTGTT	1200
	GGTTTACGGT TTCTGTGTGT GTCTTGTGTG TCTCTTGTG TTCAGACTTG GACTGATGAC	1260
	TGACGACTGT TTTTAAGTTA TGCCTCTAA AATAAGCCTA AAAATCCTGT CAGATCCCTA	1320
40	TGCTGACCAC TTCCCTTCAG ATCAACAGCT GCCCTTACTC GAGCTCAAGC TTGAAATTCT	1380
	GCAGTCGACG GTACCGCGGC CGCTAACTAA TAGCCCATTG TCCAAGGTAC GTAGCGGGGA	1440
45	TCAATTCCGC CCCCCCCCCTA ACGTTACTGG CCGAAGCCGC TTGGAATAAG GCCGGTGTGC	1500
	GTGGTCTAT ATGTTATTT CCACCATATT GCCGTCTTT GGCAATGTGA GGGCCCGGAA	1560
	ACCTGGCCCT GTCTTCTTGA CGAGCATTCC TAGGGGTCTT TCCCTCTCG CCAAAGGAAT	1620
50	GCAAGGTCTG TTGAATGTG TGAGGAAAGC AGTCCTCTG GAAGCTTCTT GAAGACAAAC	1680
	AACGTCTGTA GCGACCCTT GCAGGCAGCG GAACCCCCA CCTGGCGACA GGTGCCTCTG	1740
55	CGGCCAAAAG CCACGTGTAT AAGATACACC TGCAAAGGCG GCACAACCCC AGTGCCACGT	1800
	TGTGAGTTGG ATAGTTGTGG AAAGAGTCAA ATGGCTCTCC TCAAGCGTAT TCAACAAGGG	1860
	GCTGAAGGAT GCCCAGAAGG TACCCATTG TATGGGATCT GATCTGGGCG CTCGGTGCAC	1920
60	ATGCTTTACA TGTGTTAGT CGAGGTTAAA AAAACGTCTA GGCCCCCGA ACCACGGGA	1980
	CGTGGTTTC CTTTGAAAAA CACGATACGG GATCCACCGG TCGCCACCAT GGGTAAAGGA	2040
65	GAAGAACTTT TCACAGGAGT TGTCCCAATT CTTGTTGAAT TAGATGGTGA TGTAAATGGG	2100
	CACAAATTTT CTGTCAGTGG AGAGGGTGAA GGTGATGCAA CATACGGAAA ACTTACCCCTT	2160

	AAATTTATTT	GCACTACTGG	AAAAC TACCT	GTTCCATGGC	CAACACTTGT	CACTACTTTC	2220
	ACTTATGGTG	TTCAATGCTT	TTCAAGATA C	CCAGATCATA	TGAAACGGCA	TGACTTTTC	2280
5	AAGAGTGCCA	TGCCCGAAGG	TTATGTACAG	GAAAGAAC TA	TATTTTCAA	AGATGACGGG	2340
	AACTACAAGA	CACGTGCTGA	AGTCAAGTT	GAAGGTGATA	CCCTGTAA	TAGAATCGAG	2400
10	TTAAAAGGTA	TTGATTTAA	AGAAGATGGA	ACATTCTG	GACACAA ATT	GGAATACAAC	2460
	TATAACTCAC	ACAATGTATA	CATCATGGCA	GACAAACAAA	AGAATGGAAC	CAAAGTTAAC	2520
	TTCAAAATTA	GACACAACAT	TGAAGATGGA	AGCGTTAAC	TAGCAGACCA	TTATCAACAA	2580
15	AATACTCCAA	TTGGCGATGG	CCCTGTCCTT	TTACCAGACA	ACCATTACCT	GTCCACACAA	2640
	TCTGCCCTT	CGAAAGATCC	CAACGAAAAG	AGAGACCACA	TGGTCCTTCT	TGAGTTGTA	2700
20	ACAGCTGCTG	GGATTACACA	TGGCATGGAT	GAAC TATA CA	AGTCCGGATC	TAGATAACTG	2760
	TATCGATGGA	TCCGAAGGCG	GGGACAGCAG	TGCAGTGGTG	GACAGAAAGC	AAGTGATCTA	2820
	GGCCAGCAGC	CTCCCTAAAG	GGACTTCAGC	CCACAAAGCC	AAACTTGTGG	CTTTAATACA	2880
25	AGCTCTGTAA	ATGGTAAAAA	AAAAAAAGTC	TACACGGACA	GCAGGTATGC	TCTTGCCACT	2940
	GTACAGAGCA	ATATA CAGAC	AAAGAGAACT	GTTGACATCT	GCAGAGAAAG	ACCTAAGATG	3000
	CTGTGGCTAA	AAGAAATCAG	ATGGCAAATC	TAACCGCCC A	GGCATCCTAA	AGAGCAATGA	3060
30	TCCTGACAGT	CTGAAGACTA	TCAAGTTATA	GACAAATTAA	GA CTGGTAA	AAAACCCCTG	3120
	TATAAAATAG	AAAAAACTGA	AAAAAGAAAA	CTAGTCCTCT	CATGAGAAGA	CAGACCTGAC	3180
35	ATCTACTGAA	AAATAGACTT	TACTGGAAAA	AATATGTGTA	TGAATACCTT	CTAGTTTTG	3240
	TGAACGTTCT	CAAGATGGAT	AAAAGCTTT	CCTTGTAAA	CGAGACTGAT	CAGATAGTCA	3300
	TCAAGAAGAT	TGTTAAAGAA	AATTTCAA	GGTCGGAGT	GCCAAAAGCA	ATAGTGTCA G	3360
40	ATAATGGTCC	TGCCCTTGTT	GCCCAGGTAA	GTCAGGGTGT	GGCCAAGTAT	TTAGAGGTCA	3420
	AATGAAAATT	CCATTGTGTG	TACAGACCTC	AGAGCTCAGG	AAAGATAAAA	AAGAATAAAAT	3480
45	AAAAC TCTAA	ACAGACCTTG	ACAAAATTAA	TCCTAGAGAC	TGGCACAGAC	TTACTGGTA	3540
	CTCCTTCCCC	TTGCCCTATT	TAGAACTGAG	AATACTCCCT	CTTGATT CGG	TTTTACTCTT	3600
50	TTTAAGATCC	TTTATGGGGC	TCCTATGCCA	TC ACTGTCTT	AAATGATGTG	TTTAAACCTA	3660
	TGTTGTTATA	ATAATGATCT	ATATGTTAAG	TTAAAAGGCT	TGCAGGTGGT	GCAGAAAGAA	3720
	GTCTGGTCAC	AACTGGCTAC	AGTGAACAAG	CTGGGTACCC	CAAGGACATC	TTACCAGTTC	3780
55	CAGCCAGAGA	TCTGATCTAC	GATCCCCGGG	TCGACCCGGG	TCGACCCGT	GGAATGTGTG	3840
	TCAGTTAGGG	TGTGGAAAGT	CCCCCAGGCTC	CCCAGCAGGC	AGAAGTATGC	AAAGCATGCA	3900
	TCTCAATTAG	TCAGCAACCA	GGTGTGGAAA	GTCCCCAGGC	TCCCCAGCAG	GCAGAAGTAT	3960
60	GCAAAGCATG	CATCTCAATT	AGTCAGCAAC	CATAGTCCCG	CCCCTAACTC	CGCCCATCCC	4020
	GCCCCCTAACT	CCGCCAGTT	CCGCCCATTC	TCCGCCCAT	GGCTGACTAA	TTTTTTTAT	4080
	TTATGCAGAG	GCCGAGGCCG	CCTCGGCCTC	TGAGCTATT C	CAGAAGTAGT	GAGGAGGCTT	4140
65	TTTTGGAGGC	CTAGGCTTT	GCAAAAAGCT	TCACGCTGCC	GCAAGCACTC	AGGGCGCAAG	4200

	GGCTGCTAAA GGAAGCGGAA CACGTAGAAA GCCAGTCCGC AGAACCGGTG CTGACCCCCGG	4260
5	ATGAATGTCA GCTACTGGGC TATCTGGACA AGGGAAAACG CAAGCGCAA GAGAAAGCAG	4320
	GTAGCTTGCA GTGGGCTTAC ATGGCGATAG CTAGACTGGG CGGTTTATG GACAGCAAGC	4380
	GAACCGGAAT TGCCAGCTGG GGCGCCCTCT GGTAAGGTTG GGAAGCCCTG CAAAGTAAAC	4440
10	TGGATGGCTT TCTTGCCGCC AAGGATCTGA TGGCGCAGGG GATCAAGATC TGATCAAGAG	4500
	ACAGGATGAG GATCGTTTCG CATGATTGAA CAAGATGGAT TGCACGCAGG TTCTCCGGCC	4560
15	GCTTGGGTGG AGAGGCTATT CGGCTATGAC TGGGCACAAC AGACAATCGG CTGCTCTGAT	4620
	GCCGCCGTGT TCCGGCTGTC AGCGCAGGGG CGCCCGGTTTC TTTTGTCAA GACCGACCTG	4680
	TCCGGTGCCCG TGAATGAACG GCAGGACGAG GCAGCGCGGC TATCGTGGCT GGCCACGACG	4740
20	GGCGTTCCTT GCGCAGCTGT GCTCGACGTT GTCACTGAAG CGGGAAAGGGA CTGGCTGCTA	4800
	TTGGGCGAAG TGCCGGGGCA GGATCTCCTG TCATCTCACCG TTGCTCCTGC CGAGAAAGTA	4860
25	TCCATCATGG CTGATGCAAT GCGGCGGCTG CATA CGCTTG ATCCGGCTAC CTGCCCATTC	4920
	GACCACCAAG CGAAACATCG CATCGAGCGA GCACGTACTC GGATGGAAGC CGGTCTTGTC	4980
	GATCAGGATG ATCTGGACGA AGAGCATCAG GGGCTCGCGC CAGCCGAACG GTTCGCCAGG	5040
30	CTCAAGGCGC GCATGCCCGA CGGCGAGGAT CTCGTCGTGA CCCATGGCGA TGCCTGCTTG	5100
	CCGAATATCA TGGTGGAAAA TGGCCGCTTT TCTGGATTCA TCGACTGTGG CC GGCTGGGT	5160
35	GTGGCGGACC GCTATCAGGA CATAGCGTTG GCTACCCGTG ATATTGCTGA AGAGCTTGGC	5220
	GGCGAATGGG CTGACCGCTT CCTCGTGCTT TACGGTATCG CCGCTCCCGA TTCGCAGCGC	5280
	ATCGCCTTCT ATCGCCTTCT TGACGAGTTC TTCTGAGCGG GACTCTGGGG TTCGAAATGA	5340
40	CCGACCAAGC GACGCCAAC CTGCCATCAC GAGATTTCGA TTCCACCGCC GCCTTCTATG	5400
	AAAGGTTGGG CTTCGGAATC GTTTCCGGG ACGGAATTG TAATCTGCTG CTTGCAAACA	5460
45	AAAAAACAC CGCTACCAGC GGTGGTTGT TTGCCGGATC AAGAGCTACC AACTCTTTT	5520
	CCGAAGGTAA CTGGCTTCAG CAGAGCGCAG ATACCAAATA CTGTCCTTCT AGTGTAGCCG	5580
	TAGTTAGGCC ACCACTTCAA GAACTCTGTA GCACCGCCTA CATACTCGC TCTGCTAATC	5640
50	CTGTTACCAG TGGCTGCTGC CAGTGGCGAT AAGTCGTGTC TTACCGGGTT GGACTCAAGA	5700
	CGATAGTTAC CGGATAAGGC GCAGCGGTG GGCTGAACGG GGGGTTCGTG CACACAGCCC	5760
55	AGCTTGGAGC GAACGACCTA CACCGAACTG AGATACCTAC AGCGTGAGCA TTGAGAAAGC	5820
	GCCACGCTTC CGCGAAGGGAG AAAGGCGGAC AGGTATCCGG TAAGCGGCAG GGTCGGAACA	5880
	GGAGAGCGCA CGAGGGAGCT TCCAGGGGAA AACGCCTGGT ATCTTTATAG TCCTGTCGGG	5940
60	TTTCGCCACC TCTGACTTGA GCGTCGATTT TTGTGATGCT CGTCAGGGGG GCGGAGCCTA	6000
	TGGAAAAACG CCAGCAACGC CGAGATGCGC CGCCTCGAGT ACACCTGCGT CATGCTGAGA	6060
	CCCTCAAGCC TCACTAAAAG GGTCCCTGCC TAGTTCTGTT TACTAATCTG CCTTATTCTG	6120
65	TTTTTGTTCATGTTAAAG ATAGAGTAAA TGCAGTATTC TCCACATAGA GATATAGACT	6180

TCTGAAATTC TAAGATTAGA ATTATTACCA AGAAGAAGTG GGGAA

6225-

## (2) INFORMATION FOR SEQ ID NO:2:

- 5           (i) SEQUENCE CHARACTERISTICS:  
              (A) LENGTH: 487 base pairs  
              (B) TYPE: nucleic acid  
              (C) STRANDEDNESS: single  
              (D) TOPOLOGY: linear  
10           (ii) MOLECULE TYPE: DNA (genomic)

- 15           (xi) SEQUENCE DESCRIPTION: SEQ ID NO:2:

CCTCCCATCT	AGAGGTTGTT	CTCGGAACAC	TCTAAACTT	TTCACCCCAA	AACTCCTCAC	60
20 CCTAAAGTTC	GAAAAAACTG	TTCCAAGAAC	ATTTTGAGA	TAAAGGCCTC	CTAGAACAAAC	120
CTCAAAATGA	CATTGCCAAA	TGATAAGACA	TGACTCCTTA	GTTACGTAGG	TTCCTTGATA	180
25 GGACATGACT	CCTTAGTTAC	GTAGGTTCT	TGATAGGACA	TGACTCCTTA	GTTACGTAGA	240
TTCCTTGTT	AGAACTCCCT	AGTGATGTAA	ACTTGTACTT	TCCCTGCCCA	GTTCTCCCCC	300
30 CTGTGCTAAG	GTGTATGAGT	TTCGACCCCA	GAGCTCTGTG	TGCTTCCATG	TTGCTGCTTT	420
ATTCGACCC	CAGAGCTCTG	GTCTGTGTGC	TTTCATGTGC	CTGCTTTATT	AAATCTTGCC	480
TTCTACA						487

- 35           (2) INFORMATION FOR SEQ ID NO:3:

40           (i) SEQUENCE CHARACTERISTICS:  
              (A) LENGTH: 366 base pairs  
              (B) TYPE: nucleic acid  
              (C) STRANDEDNESS: single  
              (D) TOPOLOGY: linear  
45           (ii) MOLECULE TYPE: DNA (genomic)

- 50           (xi) SEQUENCE DESCRIPTION: SEQ ID NO:3:

CCTCCCATCT	AGAAAACATT	TTTGAGATAA	AGGCTTCCTG	GAACAACCTC	AAAATGAACC	60
AGGTACTCCT	TAGTTACGTA	GGTCCTTGA	TAGGACATGA	CTCCTTAGTT	ACATAGATTC	120
55 CTTTGGCAGA	ACTCCCTAGT	GATGTAAACT	TGTACTTTCC	CTGCCAGTT	CTCCCCCTTT	180
GAGTTTACT	ATATAAGCCT	GTGAAAATT	TTGGCTGACC	GTCGAGACTC	CTCTACCTG	240
60 TGCTAAGGTG	TATGAGTTTC	GACCCCAGAG	CTCTGTGTGC	TTCCATGTTG	CTGCTTTATT	300
TCGACCCAG	AGCTCTGGTC	TGTGTGCTT	CATGTTGCTG	CCTTATTAAA	TCTTGCCTTC	360
TACATT						366

- 65           (2) INFORMATION FOR SEQ ID NO:4:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 304 base pairs
  - (B) TYPE: nucleic acid
  - (C) STRANDEDNESS: single
  - (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

10 (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 4:

CCTCCCATCT AGAGATTGTT CCCAGAACAC TCCTGAACTC TTCACCCCCAG AATGCATGCC 60  
TGAACCTCTC ACCCTAGAGT TCGAACCCCTC CCAACTAAAG ACTGTTCCAA GAACATTTT 120  
GAGATAAGGG CCTCCTGGAA CAACCTCAGA ATGAACCGGG TACATTGCCA AATAATAGGA 180  
CATGACCCCT TAGTTACGTA AAATCCCTTG GCAGAACCCC TTGTCCCTTG GCAGAACCCC 240  
TTAGTTATGT AAACTTGTAC TTTCCCTTACCC CGGCTCTCCCC CCCTTGAGTT TTTCCATATAT 300  
AAGC 304

25 (2) INFORMATION FOR SEQ ID NO:5:

- (i) SEQUENCE CHARACTERISTICS:

  - (A) LENGTH: 304 base pairs
  - (B) TYPE: nucleic acid
  - (C) STRANDEDNESS: single
  - (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

35

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:5:

40	CCTCCCCATCT AGAGAGTGTT CCCAGAACAC TCCTGAACTC TTCACCCAG AATGCATTCC	60
	TGAACTCCTC ACCCTAGAGT TCGAACCCCTC CCAACTAAAG ACTGTTCCAA GAACATTTT	120
45	GAGATAAGGG CCTCCTGGAA CAACCTCAGA ATGAACCGGA TACATTGCCA AATAATAGGA	180
	CATGACCCCT TAGTTACGTA GAATCCCTTG GCAGAACCCC TTGTCCCTTG GCAGAACCCC	240
	TTAGTTATGT AAACTTGTAC TTTCCCTACC CCGCTCTCCC CCCTTGAGTT TTTCCCTATAT	300
50	AAGC	304

(2) INFORMATION FOR SEQ ID NO:6:

- (i) SEQUENCE CHARACTERISTICS:

  - (A) LENGTH: 304 base pairs
  - (B) TYPE: nucleic acid
  - (C) STRANDEDNESS: single
  - (D) TOPOLOGY: linear

60 (ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:6:

65 CCTCCCCATCT AGAGAGTGT CCCAGAACAC TCCTGAACTC TTCACCCAG AATGCATTCC 60

	TGAACTCCTC ATCCTAGAGT TCGAACCCCTC CCAACTAAAG ACTGTTCCAA GAACATTTT	120
5	GAGATAAGGG CCTCCTGGAA CAACCTCAGA ATGAACCTGG TACATTGCCA AATAATAGGA	180
	CATGACCCTT TAGTTACGTA GAATCCCTTG GCAGAACCCC TTGTCCCTTG GCAGAACCCC	240
	TTAGTTATGC AAACTTGTAC TTTCTCTGCC CCGCTCTCCC CCCTTGAGTT TTTCCTATAT	300
10	AAGC	304

## (2) INFORMATION FOR SEQ ID NO:7:

- 15 (i) SEQUENCE CHARACTERISTICS:  
 (A) LENGTH: 304 base pairs  
 (B) TYPE: nucleic acid  
 (C) STRANDEDNESS: single  
 (D) TOPOLOGY: linear
- 20 (ii) MOLECULE TYPE: DNA (genomic)

## 25 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:7:

	CCTCCCATCT AGAGAGTGT TCCAGAACAC TCCTGAACTC TTCACCTCAA AATGCATTCC	60
30	TGAACTCCTC ACCCTAGAGT TCGAACCCCTC CCAACTAAAG ACTGTTCCAA GAACATTTT	120
	GAGATAAGGG CCTCCTGGAA CAACCTCAGA ATGAACCGAG TACATTGCCA AATAATAGGA	180
	CATGACCCTT TAGTTACGTA GAATCCCTTG GCAGAACCCC TTGTCCCTTG GCAGAACCCC	240
35	TTAGTTATGC AAACTTGTAC TTTCTCTGCC CCGCTCTCCC CCCTTGAGTT TTTCCTATAT	300
	AAGC	304

## 40 (2) INFORMATION FOR SEQ ID NO:8:

- 45 (i) SEQUENCE CHARACTERISTICS:  
 (A) LENGTH: 305 base pairs  
 (B) TYPE: nucleic acid  
 (C) STRANDEDNESS: single  
 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

## 50 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:8:

55	CCTCCCATCT AGAGATTGTT CCCAGAACAC TCCTGAACTC TTCACCCAG AATGCATTCC	60
	TGAACTCCTC ACCCTAGAGT TCGAACCCCTC CCAACTAAAG ACTGTTCCAA GAACATTTT	120
	GAGATAAGGG CCTCCTGGAA CAACCTCAGA ATGAACCGGA TACATTGCCA AATAATAGGA	180
60	CATGACCCCT TAGTTACGTA GAATTCCCTT GGCAGAACCC CTTGTCCCTT GGCAGAACCC	240
	CTTAGTTATG CAAACTTGTA CTTCCTGC CCCGCTCTCC CCCCTTGAGG TTTTCCTATA	300
65	TAAGC	305

## (2) INFORMATION FOR SEQ ID NO:9:

- (i) SEQUENCE CHARACTERISTICS:

  - (A) LENGTH: 305 base pairs
  - (B) TYPE: nucleic acid
  - (C) STRANDEDNESS: single
  - (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

10

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:9:

15	CCTCCCATCT AGAGAGTGTT CCCAGAACAC TCCTGAACTC TTCACCCCAG AATGCATTCC	60
	TGAACCCCTC ACCCTAGAGT TCGAACCCCTC CCAACTAAAG ACTGTTCCAA GAACATTTT	120
20	GAGATAAGGG CCTCCTGGAA CAACCTCAGA ATGAACCAGG TACATTGCCA ATAATAGGA	180
	CATGACCCCT TAGTTACGTA GAATTCCCTT GGCAGAACCC CTTGTCCCTT GGCAGAACCC	240
	CTTAGTTATG CGAACTTGTA CTTTCCCTGC CCCGCTCTCC CCCCTTGAGT TTTTCCTATA	300
25	TAAGC	305

(2) INFORMATION FOR SEQ ID NO:10:

- (i) SEQUENCE CHARACTERISTICS:

  - (A) LENGTH: 306 base pairs
  - (B) TYPE: nucleic acid
  - (C) STRANDEDNESS: single
  - (D) TOPOLOGY: linear

35 (ii) MOLECULE TYPE: DNA (genomic)

40 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:10:

CCCTCCCATC	TAGAGAGTGT	TCCCAGAACCA	CTCCTGAACCT	CTTCATCCCCA	GAATGCATTC	60
CTGAACTCCT	CACCCTATAG	TTCGAACCCCT	CCCAAACCTAAA	GACTGTTCCA	AGAACATTTC	120
TGAGATAAGG	GCCTCCTGGA	ACAACCTCAG	AATGAACCGG	GTACATTGCC	AAATAATAGG	180
ACATGACCCCC	TTAGTTACGT	AGAATTCCCT	TGGCAGAACCC	CCTTGTGCGCT	TGGCAGAACCC	240
CCTTAGTTAT	GTAAACTTGT	ACTTTCCCTG	CCCCGCTCTC	CCCCCTTGAG	TTTTTACTAT	300
ATAAGC						306

(2) INFORMATION FOR SEQ ID NO:11:

- (i) SEQUENCE CHARACTERISTICS:

  - (A) LENGTH: 305 base pairs
  - (B) TYPE: nucleic acid
  - (C) STRANDEDNESS: single
  - (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

65

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:11:

	CCTCCCATCT	AGAGAGTGTT	CCCAAAACAC	TCCTGAACTC	TTCACCCCCAG	AATGCATTCC	60
5	TGAACTCCTC	ACCCTAAAGT	TCAAACCCCTC	CCAACTAAAG	ACTGTTCCAA	GAACATTTT	120
	GAGATAAGGG	CCTCCTGGAA	CAACCTCAGA	ATGAACCGGG	TACATTGCCA	AATAATAGGA	180
10	CATGACCCCT	TAGTTACACA	GAATTCCCTT	GGCAAAACCC	CTTGTCCCTT	GGCAGAACCC	240
	CTTAGTTATG	CAAACTTGTA	CTTCCCCTGC	CCAGCTCTCC	CCCCTTGAGT	TTTTCCTATA	300
	TAAGC						305

15 (2) INFORMATION FOR SEQ ID NO:12:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 304 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

25

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:12:

30	CCTCCCATCT AGAGAGTGTT CCCAGAACAC TCCTGAACTC TTCACCCAG AATGCATTCC	60
	TGAACTCCTC ACCCTAGAGT TTGAACCCTC CCAACTAAAG ACTGTTCCAA GAACATCTT	120
35	GAGATAAGGG CCTCCTGGAA CAACCTCAGA ATGAACCGGG TACATTGCCA ATAATAGGA	180
	CATGACCCCT TAGTTACGTA GAATTCCCTT GGCAGAACCC CTTGTCGCTT GGCAGAACCC	240
	CTTAGTTATG CAAACTTGTA CTTCCTGC CCCGCTCTCC CCCTTGAGTT TTTCCTATAT	300
40	AAGC	304

(2) INFORMATION FOR SEO ID NO:13:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 303 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

50 (ii) MOLECULE TYPE: DNA (genomic)

55 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:13:

	CCTCCCATCT	AGAGAGTGT	CCCAGAACAC	TCCTAAACTC	TTCACCCAG	AATGCATTCC	60
	TGAACTCCTC	ACCCTAGAGT	TCGAACCCTT	CCAACTAAAG	ACTGTTCCAA	GAACATTTT	120
60	GAGATAAGGG	CCTCCTGGAA	CAACCTCAAA	ATGAACCGGG	TACATTGCCA	AATGATAGGA	180
	CATGACCCCT	TAGTTACGTA	GATTCCCTTG	GCAGAACCCCC	TTGTCCCTTG	GCAGAACCCCC	240
65	CTAGTGATGT	AAACTTGTAC	TTTCCCTGCC	CAGCTCTCCC	CCCTTGAGTT	TTCCTATATA	300
	AGC						303

## (2) INFORMATION FOR SEQ ID NO:14:

## (i) SEQUENCE CHARACTERISTICS:

- 5           (A) LENGTH: 8657 base pairs  
           (B) TYPE: nucleic acid  
           (C) STRANDEDNESS: single  
           (D) TOPOLOGY: linear

10           (ii) MOLECULE TYPE: DNA (genomic)

15           (xi) SEQUENCE DESCRIPTION: SEQ ID NO:14:

TGAAGAATAA AAAATTACTG GCCTCTTGTG AGAACATGAA CTTTCACCTC GGAGCCCACC	60
CCCTCCCACAT TGGAAAACAT ACTTGAGAAA AACATTTCT GGAACAACCA CAGAATGTTT	120
20           CAACAGGCCA GATGTATTGC CAAACACAGG ATATGACTCT TTGGTTGAGT AAATTTGTGG	180
TTGTTAAACT TCCCCTATTTC CCTCCCCATT CCCCTCCCA GTTGTGGTT TTTTCCTTA	240
25           AAAGCTTGTG AAAAATTGA GTCGTCGTCG AGACTCCTCT ACCCTGTGCA AAGGTGTATG	300
AGTTTCGACC CCAGAGCTCT GTGTGCTTTC TGTTGCTGCT TTATTCGAC CCCAGAGCTC	360
30           TGGTCTGTGT GCTTCATGT CGCTGCTTTA TTAAATCTTA CCTTCTACAT TTTATGTATG	420
GTCTCAGTGT CTTCTGGGT ACGCGGCTGT CCCGGGACTT GAGTGTCTGA GTGAGGGTCT	480
TCCCTCGAGG GTCTTCATT TGGTACATGG GCCGGGAATT CGAGAATCTT TCATTTGGTG	540
35           CATTGCCCG GAATTGAAATCTTCATT TGGTGCATTG GCCGGGAAAC AGCGCGACCA	600
CCCAGAGGTC CTAGACCCAC TTAGAGGTAA GATTCTTGT TCTGTTTG TGATGTCT	660
40           GTGTTCTGAT GTCTGTGTT TCAGTGTCTAAG TCTGGTGCAGA TCGCAGTTTC AGTTTGCGG	720
ACGCTCAGTG AGACCGCGCT CCGAGAGGGA GTGCGGGGTG GATAAGGATA GACGTGTCCA	780
GGTGTCCACC GTCCGTTCGC CCTGGGAGAC GTCCCAGGAG GAACAGGGGA GGATCAGGGA	840
45           CGCCTGGTGG ACCCCATTGA AGGCCAAGAG ACCATTGGG GTTGCAGAT CGTGGGTTCG	900
AGTCCCACCT CGTGCCAGT TGCGAGATCG TGGGTTCGAG TOCCACCTCG TGTTTGTG	960
50           CGAGATCGTG GGTCGAGTC CCACCTCGCG TCTGGTCACG GGATCGTGGG TTCGAGTCCC	1020
ACCTCGTGT TTGTTGCGAG ATCGTGGGTT CGAGTCCCAC CTCGCGTCTG GTCACGGGAT	1080
CGTGGGTTCG AGTCCCACCT CGTGCAAGGG GTCTCAATTG GCCGGCCTTA GAGAGGCCAT	1140
55           CTGATTCTTC TGGTTCTCT TTTTGTCTTA GTCTCGTGT CGCTCTGTT GTGACTACTG	1200
TTTTCTAAA AATGGGACAA TCTGTGTCCA CTCCCCTTTC TCTGACTCTG GTTCTGTCGC	1260
60           TTGGTAATTG TGTTTGTGTTA CGTTTGTGTT TGTGAGTCGT CTATGTTGTC TGTTACTATC	1320
TTGTTTTGT TTGTTGTTA CGGTTCTGT GTGTGTCTTG TGTGTCCTT TGTGTTCAGA	1380
CTTGGACTGA TGACTGACGA CTGTTTTAA GTTATGCCTT CTAAAATAAG CCTAAAAATC	1440
65           CTGTCAGATC CCTATGCTGA CCACCTCCTT TCAGATCAAC AGCTGCCCTT ACgtatcgat	1500

	GGATCCCTCG ACTAACTAAT AGCCCATTCT CCAAGGTCGA GCGGGATCAA TTCCGCC	1560
	CCCCTAACGT TACTGGCCGA AGCCGTTGG AATAAGGCCG GTGTGCGTT GTCTATATGT	1620
5	TATTTCCAC CATATTGCCG TCTTTGGCA ATGTGAGGGC CCGGAAACCT GGCCCTGTCT	1680
	TCTTGACGAG CATTCTAGG GGTCTTCCC CTCTGCCAA AGGAATGCAA GGTCTGTTGA	1740
10	ATGTCGTGAA GGAAGCAGTT CCTCTGGAAG CTTCTGAAG ACAAAACAACG TCTGTAGCGA	1800
	CCCTTGCAG GCAGCGGAAC CCCCCACCTG GCGACAGGTG CCTCTGCGC CAAAAGCCAC	1860
	GTGTATAAGA TACACCTGCA AAGGCGGCAC AACCCCAGTG CCACGTTGTG AGTTGGATAG	1920
15	TTGTGGAAAG AGTCAAATGG CTCTCCTCAA GCGTATTCAA CAAGGGGCTG AAGGATGCC	1980
	AGAAGGTACC CCATTGTATG GGATCTGATC TGGGGCCTCG GTGCACATGC TTTACATGTG	2040
20	TTTAGTCGAG GTAAAAAAA CGTCTAGGCC CCCCCAACCA CGGGGACGTG GTTTCC	2100
	GAAAAACACG ATAATAATCA TGGGCGCGGA TCCCCTCGTT TTACAACGTC GTGACTGG	2160
	AAACCCCTGGC GTTACCCAAC TTAATCGCCT TGCAGCACAT CCCCCTTCG CCAGCTGGCG	2220
25	TAATAGCGAA GAGGCCCGCA CCGATGCC CCCAAACAG TTGCGCAGCC TGAATGGCGA	2280
	ATGGCGCTTT GCCTGGTTTC CGGCACCAGA AGCGGTGCCG GAAAGCTGGC TGGAGTGC	2340
	TCTTCCTGAG GCCGATACTG TCGTCGTCCC CTCAAACTGG CAGATGCACG GTTACGATGC	2400
30	GCCCCATCTAC ACCAACGTAA CCTATCCCAT TACGGTCAAT CCGCCGTTTG TTCCCACGGA	2460
	GAATCCGACG GGTTGTTACT CGCTCACATT TAATGTTGAT GAAAGCTGGC TACAGGAAGG	2520
35	CCAGACCGA ATTATTTTG ATGGCGTTAA CTCGGCGTTT CATCTGTGGT GCAACGGCG	2580
	CTGGGTCGGT TACGGCCAGG ACAGTCGTTT GCCGCTGAA TTTGACCTGA GCGCATT	2640
40	ACGCGCCGGA GAAAACCGCC TCGCGGTGAT GGTGCTGCGT TGGAGTGACG GCAGTTATCT	2700
	GGAAGATCAG GATATGTGGC GGATGAGCGG CATTTCCTGT GACGTCTCGT TGCTGCATAA	2760
	ACCGACTACA CAAATCAGCG ATTTCCATGT TGCCACTCGC TTTAATGATG ATTCAGCCG	2820
45	CGCTGTACTG GAGGCTGAAG TTCAGATGTG CGGCGAGTTG CGTACTACC TACGGTAAC	2880
	AGTTTCTTTA TGGCAGGGTG AAACGCAGGT CGCCAGCGGC ACCGCGCCTT TCGGCGGTGA	2940
	AATTATCGAT GAGCGTGGTG GTTATGCCGA TCGCGTCACA CTACGTCTGA ACGTCGAAA	3000
50	CCCGAAACTG TGGAGCGCCG AAATCCGAA TCTCTATCGT GCGGTGGTTG AACTGCACAC	3060
	CGCCGACGGC ACGCTGATTG AAGCAGAAGC CTGCGATGTC GGTTCCGCG AGGTGCGGAT	3120
55	TGAAAATGGT CTGCTGCTGC TGAACGGCAA GCCGTTGCTG ATTGAGGCG TTAACCGTCA	3180
	CGAGCATCAT CCTCTGCATG GTCAGGTCA GGATGAGCAG ACGATGGTGC AGGATATCCT	3240
	GCTGATGAAG CAGAACAACT TTAACGCCGT GCGCTGTTG CATTATCCGA ACCATCCGCT	3300
60	GTGGTACACG CTGTGCGACC GCTACGGCCT GTATGTGGTG GATGAAGCCA ATATTGAAAC	3360
	CCACGGCATG GTGCCAATGA ATCGTCTGAC CGATGATCCG CGCTGGCTAC CGGCGATGAG	3420
	CGAACCGCGTA ACGCGAATGG TGCAGCGCGA TCGTAATCAC CCGAGTGTGA TCATCTGGTC	3480
65	GCTGGGAAT GAATCAGGCC ACGGGCCTAA TCACGACGCG CTGTATCGCT GGATCAAATC	3540

	TGTCGATCCT TCCC GCCCGGG TGCAGTATGA AGGC GGCGGA GCCGACACCA CGGCCACCGA	3600
5	TATTATTGC CCGATGTACG CGCGCGTGG A TGAAGACCAG CCCTTCCCGG CTGTGCCGAA	3660
	ATGGTCCCATC AAAAAATGGC TTTCGCTACC TGGAGAGACG CGCCCGCTGA TCCTTGCGA	3720
	ATACGCCAC GCGATGGGTA ACAGTCTTGG CGGTTTCGCT AAATACTGGC AGGCGTTCG	3780
10	TCAGTATCCC CGTTTACAGG GCGGCTTCGT CTGGGACTGG GTGGATCAGT CGCTGATTAA	3840
	ATATGATGAA AACGGCAACC CGTGGTCGGC TTACGGCGGT GATTTGGCG ATACGCCGAA	3900
15	CGATGCCAG TTCTGTATGA ACGGCTGGT CTTGCCGAC CGCACGCCGC ATCCAGCGCT	3960
	GACGGAAGCA AAACACCAGC AGCAGTTTT CCAGTTCCGT TTATCCGGGC AAACCATCGA	4020
	AGT GACCAGC GAATACTGT TCCGTATAG CGATAACGAG CTCCTGCCT GGATGGTGGC	4080
20	GCTGGATGGT AAGCCGCTGG CAAGCGGTGA AGTGCCTCTG GATGTCGCTC CACAAGGTAA	4140
	ACAGTTGATT GAACTGCCTG AACTACCGCA GCCGGAGAGC GCCGGGCAAC TCTGGCTCAC	4200
25	AGTACCGCTA GTGCAACCGA ACGCGACCGC ATGGTCAGAA GCCGGGCACA TCAGCGCCTG	4260
	GCAGCAGTGG CGTCTGGCGG AAAACCTCAG TGTGACGCTC CCCGCCGCGT CCCACGCCAT	4320
	CCCGCATCTG ACCACCAGCG AAATGGATT TTGCATCGAG CTGGGTAATA AGCGTTGGCA	4380
30	ATTTAACCGC CAGTCAGGCT TTCTTCACA GATGTGGATT GGCGATAAAA ACAACTGCT	4440
	GACGCCGCTG CGCGATCAGT TCACCCGTGC ACCGCTGGAT AACGACATTG GCGTAAGTGA	4500
35	AGCGACCCGC ATTGACCCCTA ACGCCTGGGT CGAACGCTGG AAGGCGGCGG GCCATTACCA	4560
	GGCCGAAGCA GCGTTGTTGC AGTGCACGGC AGATAACATT GCTGATGGCG TGCTGATTAC	4620
	GACCGCTCAC GCGTGGCAGC ATCAGGGGAA AACCTTATTG ATCAGCCGGA AAACCTACCG	4680
40	GATTGATGGT AGTGGTCAA TGGCGATTAC CGTTGATGTT GAAGTGGCGA GCGATAACACC	4740
	GCATCCGGCG CGGATTGGCC TGAACTGCCA GCTGGCGCAG GTAGCAGAGC GGGTAAACTG	4800
	GCTCGGATTA GGGCGCAAG AAAACTATCC CGACCGCCTT ACTGCCGCCT GTTTGACCG	4860
45	CTGGGATCTG CCATTGTCAG ACATGTATAC CCCGTACGTC TTCCCGAGCG AAAACGGTCT	4920
	GCGCTGCGGG ACGCGCGAAT TGAATTATGG CCCACACCAG TGGCGCGGCG ACTTCCAGTT	4980
50	CAACATCAGC CGCTACAGTC AACAGCAACT GATGGAAACC AGCCATCGCC ATCTGCTGCA	5040
	CGCGGAAGAA GGCACATGGC TGAATATCGA CGGTTCCAT ATGGGGATTG GTGGCGACGA	5100
55	CTCCTGGAGC CCGTCAGTAT CGGCGGAATT CCAGCTGAGC GCCGGTGCCT ACCATTACCA	5160
	GTTGGTCTGG TGTAAAAAT AATAATAACC GGGCAGGGGG GATCCGAAGG CGGGGACAGC	5220
	AGTGCAGTGG TGGACAGAAA GCAAGTGATC TAGGCCAGCA GCCTCCCTAA AGGGACTTCA	5280
60	GCCCACAAAG CCAAACCTGT GGCTTAATA CAAGCTCTGT AAATGGTAAA AAAAAAAAAG	5340
	TCTACACGGC CAGCAGGTAT GCTCTGCCA CTGTACAGAG CAATATACAG ACAAAAGAGAA	5400
	CTGTTGACAT CTGCAGAGAA AGACCTAAGA TGCTGTGGCT AAAAGAAATC AGATGGCAAA	5460
65	TCTAACCGCC CAGGCATCCT AAAGAGCAAT GATCCTGACA GTCTGAAGAC TATCAAGTTA	5520

	TAGACAAATT AAGACTGGTA AAAAAAACCC TGTATAAAAT AGTAAAAACT GAAAAAAGAA	5580
	AACTAGTCCT CTCATGAGAA GACAGACCTG ACATCTACTG AAAAATAGAC TTTACTGGAA	5640
5	AAAATATGTG TATGAATACC TTCTAGTTT TGTGAACGTT CTCAAGATGG ATAAAAGCTT	5700
	TTCCCTGTAA AACGAGACTG ATCAGATAGT CATCAAGAAG ATTGTTAAAG AAAATTTCC	5760
	AAGGTTCGGA GTGCCAAAAG CAATAGTGT AGATAATGGT CCTGCCCTTG TTGCCAGGT	5820
10	AAGTCAGGGT GTGGCCAAGT ATTTAGAGGT CAAATGAAAA TTCCATTGTG TGTACAGACC	5880
	TCAGAGCTCA GGAAAGATAA AAAAGAATAA ATAAAACTCT AAACAGACCT TGACAAAATT	5940
15	AATCCTAGAG ACTGGCACAG ACTTACTTGG TACTCCTTCC CCTTGCCCTA TTTAGAACTG	6000
	AGAATACTCC CTCTTGATTC GGTTTACTC TTTTAAGAT CCTTTATGGG GCTCCTATGC	6060
20	CATCACTGTC TTAAATGATG TGTTAAACC TATGTTGTTA TAATAATGAT CTATATGTTA	6120
	AGTTAAAAGG CTTGCAGGTG GTGCAGAAAG AAGTCTGGTC ACAACTGGCT ACAGTGAACA	6180
	AGCTGGGTAC CCCAAGGACA TCTTACCAAGT TCCAGCCAGA GATCTGATCT ACGATCCCCG	6240
25	GGTCGACCCG GGTGACCCCT GTGGAATGTG TGTCAGTTAG GGTGTGGAAA GTCCCCAGGC	6300
	TCCCCAGCAG GCAGAAAGTAT GCAAAGCATG CATCTCAATT AGTCAGCAAC CAGGTGTGGA	6360
	AAGTCCCCAG GCTCCCCAGC AGGCAGAAAGT ATGCAAAGCA TGCACTCTCAA TTAGTCAGCA	6420
30	ACCATAGTCC CGCCCCCTAAC TCCGCCCATC CCGCCCCTAA CTCCGCCAG TTCCGCCAT	6480
	TCTCCGCCCT ATGGCTGACT AATTTTTTT ATTATGCAG AGGCCGAGGC CGCCTCGGCC	6540
35	TCTGAGCTAT TCCAGAAGTA GTGAGGAGGC TTTTTGGAG GCCTAGGCTT TTGCAAAAG	6600
	CTTCACGCTG CCGCAAGCAC TCAGGGCGCA AGGGCTGCTA AAGGAAGCGG AACACGTAGA	6660
40	AAGCCAGTCC GCAGAACCGG TGCTGACCCC GGATGAATGT CAGCTACTGG GCTATCTGGA	6720
	CAAGGGAAAA CGCAAGCGCA AAGAGAAAGC AGGTAGCTTG CAGTGGCCTT ACATGGCGAT	6780
	AGCTAGACTG GCGGGTTTA TGGACAGCAA GCGAACCGGA ATTGCCAGCT GGGGCCCT	6840
45	CTGGTAAGGT TGGGAAGCCC TGCAAAGTAA ACTGGATGGC TTTCTGCCG CCAAGGATCT	6900
	GATGGCGCAG GGGATCAAGA TCTGATCAAG AGACAGGATG AGGATCGTT CGCATGATTG	6960
	AACAAGATGG ATTGCACGCA GGTTCTCCGG CCGCTTGGGT GGAGAGGCTA TTCGGCTATG	7020
50	ACTGGGCACA ACAGACAATC GGCTGCTCTG ATGCCGCCGT GTTCCGGCTG TCAGCGCAGG	7080
	GGCGCCCGGT TCTTTTGTG AAGACCGACC TGTCCGGTGC CCTGAATGAA CTGCAGGACG	7140
	AGGCAGCGCG GCTATCGTGG CTGGCCACGA CGGGCGTTCC TTGCGCAGCT GTGCTCGACG	7200
55	TTGTCACTGA AGCGGGAAAGG GACTGGCTGC TATTGGCGA AGTGCCGGGG CAGGATCTCC	7260
	TGTCATCTCA CCTTGCTCCT GCCGAGAAAG TATCCATCAT GGCTGATGCA ATGCGCGGC	7320
60	TGCATACGCT TGATCCGGCT ACCTGCCAT TCGACCACCA AGCGAAACAT CGCATCGAGC	7380
	GAGCACGTAC TCGGATGGAA GCCGGTCTTG TCGATCAGGA TGATCTGGAC GAAGAGCATC	7440
	AGGGGCTCGC GCCAGCCGAA CTGTTGCCA GGCTCAAGGC GCGCATGCC GACGGCGAGG	7500
65	ATCTCGTCGT GACCCATGGC GATGCCCTGCT TGCCGAATAT CATGGTGGAA AATGGCCGCT	7560

	TTTCTGGATT CATCGACTGT GGCCGGCTGG GTGTGGCGGA CCGCTATCAG GACATAGCGT	7620
5	TGGCTACCCG TGATATTGCT GAAGAGCTTG GCGGCGAATG GGCTGACCGC TTCTCGTGC	7680
	TTTACGGTAT CGCCGCTCCC GATTCCGAGC GCATCGCCTT CTATCGCCTT CTTGACGAGT	7740
	TCTTCTGAGC GGGACTCTGG GGTCGAAAT GACCGACCAA GCGACGCCA ACCTGCCATC	7800
10	ACGAGATTTC GATTCCACCG CCGCCTCTA TGAAAGGTG GGCTCGGAA TCGTTTCCG	7860
	GGACGGAATT CGTAATCTGC TGCTTGCAAA CAAAAAAACC ACCGCTACCA GCGGTGGTT	7920
15	GTTTGCCGGA TCAAGAGCTA CCAACTCTTT TTCCGAAGGT AACTGGCTTC AGCAGAGCGC	7980
	AGATAACAAA TACTGTCCTT CTAGTGTAGC CGTAGTTAGG CCACCACTTC AAGAACTCTG	8040
	TAGCACCGCC TACATACCTC GCTCTGCTAA TCCTGTTACC AGTGGCTGCT GCCAGTGGCG	8100
20	ATAAGTCGTG TCTTACCGGG TTGGACTCAA GACGATAGTT ACCGGATAAG GCGCAGCGGT	8160
	CGGGCTGAAC GGGGGGTTCG TGCACACAGC CCAGCTTGGA GCGAACGACC TACACCGAAC	8220
25	TGAGATAACCT ACAGCGTGAG CATTGAGAAA GCGCCACGCT TCCCGAAGGG AGAAAGGCGG	8280
	ACAGGTATCC GGTAAGCGGC AGGGTCGGAA CAGGAGAGCG CACGAGGGAG CTTCCAGGGG	8340
	GAAACGCCTG GTATCTTTAT AGTCCTGTG GGTTTCGCCA CCTCTGACTT GAGCGTCGAT	8400
30	TTTTGTGATG CTCGTCAGGG GGGCGGAGCC TATGGAAAAA CGCCAGCAAC GCCGAGATGC	8460
	GCCGCCTCGA GTACACCTGC GTCATGCTGA GACCCTCAAG CCTCACTAAA AGGGTCCCTG	8520
35	CCTAGTTCTG TTTACTAATC TGCCTTATTC TGTTTTGTT CCCATGTTAA AGATAGAGTA	8580
	AATGCAGTAT TCTCCACATA GAGATATAGA CTTCTGAAAT TCTAAGATTA GAATTATTTA	8640
	CAAGAAGAAG TGGGGAA	8657

## 40 (2) INFORMATION FOR SEQ ID NO:15:

- (i) SEQUENCE CHARACTERISTICS:
  - (A) LENGTH: 6359 base pairs
  - (B) TYPE: nucleic acid
  - (C) STRANDEDNESS: single
  - (D) TOPOLOGY: linear

50 (ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:15:

55	TGAAGAATAA AAAATTACTG GCCTCTTGTG AGAACATGAA CTTTCACCTC GGAGCCACC	60
	CCCTCCCACAT TGGAAACAT ACTTGAGAAA AACATTTCT GGAAACAACCA CAGAATGTT	120
60	CAACAGGCCA GATGTATTGC CAAACACAGG ATATGACTCT TTGGTTGAGT AAATTGTGG	180
	TTGTTAAACT TCCCCCTATTG CCTCCCCATT CCCCCCTCCCA GTTGTGGTT TTTTCCTTTA	240
	AAAGCTTGTG AAAAATTGAA GTCGTCGTG AGACTCCTCT ACCCTGTGCA AAGGTGTATG	300
65	AGTTTCGACC CCAGAGCTCT GTGTGCTTTC TGTTGCTGCT TTATTCGAC CCCAGAGCTC	360

	TGGTCTGTGT	GCTTCATGT	CGCTGCTTTA	TTAAATCTTA	CCTTCTACAT	TTTATGTATG	420
	GTCTCAGTGT	CTTCTTGGGT	ACGCGGCTGT	CCCAGGACTT	GAGTGTCTGA	GTGAGGGTCT	480
5	TCCCTCGAGG	GTCTTCATT	TGGTACATGG	GCCGGGAATT	CGAGAACATTT	TCATTTGGTG	540
	CATTGGCCGG	GAATTCGAAA	ATCTTCATT	TGGTGCATTG	GCCGGGAAAC	AGCGCGACCA	600
10	CCCAGAGGTC	CTAGACCCAC	TTAGAGGTAA	GATTCTTGT	TCTGTTTGG	TCTGATGTCT	660
	GTGTTCTGAT	GTCTGTGTT	TGTTTCTAAG	TCTGGTGCAG	TCGCAGTTTC	AGTTTGCAG	720
	ACGCTCAGTG	AGACCGCGCT	CCGAGAGGGA	GTGCGGGGTG	GATAAGGATA	GACGTGTCCA	780
15	GGTGTCCACC	GTCCGTTCGC	CCTGGGAGAC	GTCCCAGGAG	GAACAGGGGA	GGATCAGGGGA	840
	CGCCTGGTGG	ACCCCTTGA	AGGCCAAGAG	ACCATTGGG	GTTGCGAGAT	CGTGGGTTCG	900
20	AGTCCCACCT	CGTGCCCAGT	TGCGAGATCG	TGGGTTCCAG	TCCCACCTCG	TGTTTGTG	960
	CGAGATCGTG	GGTCGAGTC	CCACCTCGCG	TCTGGTCACG	GGATCGTGGG	TTCGAGTCCC	1020
	ACCTCGTGT	TTGTTGCGAG	ATCGTGGGTT	CGAGTCCCAC	CTCGCGTCTG	GTCACGGGAT	1080
25	CGTGGGTTCG	AGTCCCACCT	CGTGCAGAGG	GTCTCAATTG	GCCGGCCTTA	GAGAGGCCAT	1140
	CTGATTCTTC	TGGTTTCTCT	TTTTGTCTTA	GTCTCGTGT	CGCTCTTGT	GTAAGTACTG	1200
30	TTTTCTAAA	AATGGGACAA	TCTGTGTCCA	CTCCCCCTTC	TCTGACTCTG	GTTCTGTGCG	1260
	TTGGTAATT	TGTTTGTGTTA	CGTTTGTGTT	TGTGAGTCGT	CTATGTTGTC	TGTTACTATC	1320
	TTGTTTGT	TTGTTGTTA	CGGTTCTGT	GTGTGTCTTG	TGTGTCTCTT	TGTGTTCAAG	1380
35	CTTGGACTGA	TGACTGACGA	CTGTTTTAA	GTTATGCCTT	CTAAAATAAG	CCTAAAAATC	1440
	CTGTCAGATC	CCTATGCTGA	CCACTTCCTT	TCAGATCAAC	AGCTGCCCTT	ACTCGAGCTC	1500
40	AAGCTTCGAA	TTCTGCAGTC	GACGGTACCG	CGGCCGCTAA	CTAATAGCCC	ATTCTCCAAG	1560
	GTACGTAGCG	GGGATCAATT	CCGCCCCCCC	CCTAACGTTA	CTGGCCGAAG	CCGCTTGGAA	1620
	TAAGGCCGGT	GTGCGTTGT	CTATATGTTA	TTTCCACCA	TATTGCCGT	TTTGGCAAT	1680
45	GTGAGGGCCC	GGAAACCTGG	CCCTGTCTTC	TTGACGAGCA	TTCCTAGGGG	TCTTCCCCT	1740
	CTCGCCAAAG	GAATGCAAGG	TCTGTTGAAT	GTCGTGAAGG	AAGCAGTTCC	TCTGGAAGCT	1800
	TCTTGAAGAC	AAACAACGTC	TGTAGCGACC	CTTGCAGGC	AGCGGAACCC	CCCACCTGGC	1860
50	GACAGGTGCC	TCTGCGGCCA	AAAGCCACGT	GTATAAGATA	CACCTGCAA	GGCGGCACAA	1920
	CCCCAGTGCC	ACGTTGTGAG	TTGGATAGTT	GTGGAAAGAG	TCAAATGGCT	CTCCTCAAGC	1980
55	GTATTCAACA	AGGGGCTGAA	GGATGCCAG	AAGGTACCCC	ATTGTATGGG	ATCTGATCTG	2040
	GGGCCTCGGT	GCACATGCTT	TACATGTTGTT	TAGTCGAGGT	TAACAAACG	TCTAGGCC	2100
	CCGAACCACG	GGGACGTGGT	TTTCCTTGA	AAAACACGAT	ACGGGATCCA	CCGGTCGCCA	2160
60	CCATGGGTAA	AGGAGAAGAA	CTTTCACAG	GAGTTGTCCC	AATTCTTGT	GAATTAGATG	2220
	GTGATGTTAA	TGGGCACAAA	TTTCTGTCA	GTGGAGAGGG	TGAAGGTGAT	GCAACATACG	2280
	GAAAACCTTAC	CCTTAAATT	ATTTGCACTA	CTGGAAAAC	ACCTGTTCCA	TGGCCAACAC	2340
65	TTGTCACTAC	TTTCACTTAT	GGTGTCAAT	GCTTTCAAG	ATACCCAGAT	CATATGAAAC	2400

	GGCATGACTT TTTCAAGAGT GCCATGCCCG AAGGTTATGT ACAGGAAAGA ACTATATTTT	2460
5	TCAAAGATGA CGGAACTAC AAGACACGTG CTGAAGTCAA GTTGAAAGGT GATACCCTG	2520
	TTAATAGAAT CGAGTTAAAAA GGTATTGATT TTAAAGAAGA TGAAACATT CTTGGACACA	2580
	AATTGGAATA CAACTATAAC TCACACAATG TATACATCAT GGCAGACAAA CAAAAGAATG	2640
10	GAACCAAAGT TAACTTCAAA ATTAGACACA ACATTGAAGA TGGAAGCGTT CAACTAGCAG	2700
	ACCATTATCA ACAAAATACT CCAATTGGCG ATGGCCCTGT CCTTTACCA GACAACCATT	2760
15	ACCTGTCCAC ACAATCTGCC CTTTCGAAAG ATCCCACGA AAAGAGAGAC CACATGGTCC	2820
	TTCTTGAGTT TGAAACAGCT GCTGGGATTA CACATGGCAT GGATGAACTA TACAAGTCCG	2880
	GATCTAGATA ACTGTATCGA TGGATCCGAA GGCGGGGACA GCAGTGCAGT GGTGGACAGA	2940
20	AAGCAAGTGA TCTAGGCCAG CAGCCTCCCT AAAGGGACTT CAGCCCACAA AGCCAAACTT	3000
	GTGGCTTAA TACAAGCTCT GTAAATGGTA AAAAAAAAAGTCTACACG GACAGCAGGT	3060
25	ATGCTCTTGC CACTGTACAG AGCAATATAC AGACAAAGAG AACTGTTGAC ATCTGCAGAG	3120
	AAAGACCTAA GATGCTGTGG CTAAAAGAAA TCAGATGGCA AATCTAACCG CCCAGGCATC	3180
	CTAAAGAGCA ATGATCCTGA CAGTCTGAAG ACTATCAAGT TATAGACAAA TTAAGACTGG	3240
30	TAAAAAAAAC CCTGTATAAA ATAGTAAAAA CTGAAAAAAG AAAACTAGTC CTCTCATGAG	3300
	AAGACAGACC TGACATCTAC TGAAAATAG ACTTTACTGG AAAAAATATG TGTATGAATA	3360
35	CCTTCTAGTT TTTGTGAACG TTCTCAAGAT GGATAAAAGC TTTTCCTTGT AAAACGAGAC	3420
	TGATCAGATA GTCATCAAGA AGATTGTTAA AGAAAATTT CCAAGGTTCG GAGTGCCAAA	3480
	AGCAATAGTG TCAGATAATG GTCCTGCCCT TGTTGCCAG GTAAGTCAGG GTGTGCCAA	3540
40	GTATTTAGAG GTCAAATGAA AATTCCATTG TGTGTACAGA CCTCAGAGCT CAGGAAAGAT	3600
	AAAAAAGAAT AAATAAAACT CTAAACAGAC CTTGACAAAAA TTAATCCTAG AGACTGGCAC	3660
45	AGACTTACTT GGTACTCCTT CCCCTGCCCT TATTTAGAAC TGAGAATACT CCCTCTTGAT	3720
	TCGGTTTAC TCTTTTAAG ATCCTTATG GGGCTCCTAT GCCATCACTG TCTTAAATGA	3780
	TGTGTTAAA CCTATGTTGT TATAATAATG ATCTATATGT TAAGTTAAAA GGCTTGCAGG	3840
50	TGGTGCAGAA AGAAGTCTGG TCACAACTGG CTACAGTGA CAAGCTGGGT ACCCCAAGGA	3900
	CATCTTACCA GTTCCAGCCA GAGATCTGAT CTACGATCCC CGGGTCGACC CGGGTCGACC	3960
	CTGTGGAATG TGTGTCAAGT AGGGTGTGGA AAGTCCCCAG GCTCCCCAGC AGGCAGAAAGT	4020
55	ATGCAAAGCA TGCATCTCAA TTAGTCAGCA ACCAGGTGTG GAAAGTCCCC AGGCTCCCCA	4080
	GCAGGCAGAA GTATGCAAAG CATGCATCTC AATTAGTCAG CAACCATAGT CCCGCCCTA	4140
60	ACTCCGCCCA TCCCGCCCT AACTCCGCC AGTTCCGCC ATTCTCCGCC CCATGGCTGA	4200
	CTAATTTTT TTATTTATGC AGAGGCCGAG GCCGCCTCGG CCTCTGAGCT ATTCCAGAAG	4260
	TAGTGAGGAG GCTTTTTGG AGGCCTAGGC TTTTGCAAAA AGCTTCACGC TGCCGCAAGC	4320
65	ACTCAGGGCG CAAGGGCTGC TAAAGGAAGC GGAACACGTA GAAAGCCAGT CCGCAGAAC	4380

	GGTGCTGACC CCGGATGAAT GTCAGCTACT GGGCTATCTG GACAAGGGAA AACGCAAGCG	4440
	CAAAGAGAAA GCAGGTAGCT TGCAGTGGGC TTACATGGCG ATAGCTAGAC TGGGCGGTTT	4500
5	TATGGACAGC AAGCGAACCG GAATTGCCAG CTGGGGCGCC CTCTGGTAAG GTTGGGAAGC	4560
	CCTGCAAAGT AAACTGGATG GCTTCTTGC CGCCAAGGAT CTGATGGCGC AGGGGATCAA	4620
10	GATCTGATCA AGAGACAGGA TGAGGATCGT TTGCGATGAT TGAACAAGAT GGATTGCACG	4680
	CAGGTTCTCC GGCGCTTGG GTGGAGAGGC TATTGGCTA TGACTGGCA CAACAGACAA	4740
	TCGGCTGCTC TGATGCCGCC GTGTTCCGGC TGTCAGCGCA GGGCGCCCG GTTCTTTTG	4800
15	TCAAGACCGA CCTGTCCGGT GCCCTGAATG AACTGCAGGA CGAGGCAGCG CGGCTATCGT	4860
	GGCTGGCAC GACGGCGTT CCTTGCGCAG CTGTGCTCGA CGTTGTCACT GAAGCGGGAA	4920
20	GGGACTGGCT GCTATTGGGC GAAGTGCCGG GGCAGGATCT CCTGTCATCT CACCTGCTC	4980
	CTGCCGAGAA AGTATCCATC ATGGCTGATG CAATGCGCG GCTGCATAACG CTTGATCCGG	5040
	CTACCTGCC ATTGACACAC CAAGCGAAAC ATCGCATCGA GCGAGCACGT ACTCGGATGG	5100
25	AAGCCGGTCT TGTCGATCAG GATGATCTGG ACGAAGAGCA TCAGGGGCTC GCGCCAGCCG	5160
	AACTGTTCGC CAGGCTCAAG GCGCCATGC CCGACGGCGA GGATCTCGTC GTGACCCATG	5220
	GCGATGCCTG CTTGCCGAAT ATCATGGTGG AAAATGGCCG CTTTCTGGA TTCATCGACT	5280
30	GTGGCCGGCT GGGTGTGGCG GACCGCTATC AGGACATAGC GTTGGCTACC CGTGATATTG	5340
	CTGAAGAGCT TGGCGCGAA TGGGCTGACC GCTTCCTCGT GCTTACGGT ATGCCGCTC	5400
35	CCGATTGCA GCGCATCGCC TTCTATGCC TTCTTGACGA GTTCTTCTGA GCGGGACTCT	5460
	GGGGTTGAA ATGACCGACC AAGCGACGCC CAACCTGCCA TCACGAGATT TCGATTCCAC	5520
	CGCCGCCCTTC TATGAAAGGT TGGGCTTCGG AATCGTTTC CGGGACGGAA TTCGTAATCT	5580
40	GCTGCTTGCA AACAAAAAAA CCACCGCTAC CAGCGGTGGT TTGTTGCCG GATCAAGAGC	5640
	TACCAACTCT TTTCCGAAG GTAAGTGGCT TCAGCAGAGC GCAGATAACCA AATACTGTCC	5700
45	TTCTAGTGTGTA GCCGTAGTTA GGCCACCACT TCAAGAACTC TGTAGCACCG CCTACATACC	5760
	TCGCTCTGCT AATCCTGTTA CCAGTGGCTG CTGCCAGTGG CGATAAGTCG TGTCTTACCG	5820
	GGTTGGACTC AAGACGATAG TTACCGATA AGGCGCAGCG GTCGGGCTGA ACGGGGGGTT	5880
50	CGTGCACACA GCCCAGCTTG GAGCGAACGA CCTACACCGA ACTGAGATAC CTACAGCGTG	5940
	AGCATTGAGA AAGGCCACG CTTCCGAAG GGAGAAAGGC GGACAGGTAT CCGGTAAGCG	6000
	GCAGGGTCGG AACAGGAGAG CGCACGAGGG AGCTTCCAGG GGGAAACGCC TGGTATCTTT	6060
55	ATAGTCCTGT CGGGTTTCGC CACCTCTGAC TTGAGCGTCG ATTTTGTGA TGCTCGTCAG	6120
	GGGGCGGAG CCTATGGAAA AACGCCAGCA ACGCCGAGAT GCGCCGCCCTC GAGTACACCT	6180
60	GCGTCATGCT GAGACCCTCA AGCCTCACTA AAAGGGTCCC TGCTTAGTTC TGTTACTAA	6240
	TCTGCCTTAT TCTGTTTTG TTCCCATGTT AAAGATAGAG TAAATGCAGT ATTCTCCACA	6300
65	TAGAGATATA GACTTCTGAA ATTCTAAGAT TAGAATTATT TACAAGAAGA AGTGGGGAA	6359

(2) INFORMATION FOR SEQ ID NO:16:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 6891 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

10

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:16:

15	TGAAGAATAA AAAATTACTG GCCTCTTGTG AGAACATGAA CTTTCACCTC GGAGCCCACC CCCTCCCATC TGGAAAACAT ACTTGAGAAA AACATTTCT GGAACAAACCA CAGAATGTTT	60 120
20	CAACAGGCCA GATGTATTGC CAAACACAGG ATATGACTCT TTGGTTGAGT AAATTTGTGG TTGTTAAACT TCCCCTATTG CCTCCCCATT CCCCTCCCA GTTTGTGGTT TTTCTTTA	180 240
25	AAAGCTTGTG AAAAATTGAG GTCGTCGTCG AGACTCCTCT ACCCTGTGCA AAGGTGTATG AGTTTCGACC CCAGAGCTCT GTGTGCTTTC TGTTGCTGCT TTATTCGAC CCCAGAGCTC	300 360
30	TGGTCTGTG GCTTCATGT CGCTGCTTTA TTAAATCTTA CCTTCTACAT TTTATGTATG GTCTCAGTGT CTTCTGGGT ACGCGGCTGT CCCGGGACTT GAGTGTCTGA GTGAGGGTCT	420 480
35	TCCCTCGAGG GTCTTCATT TGGTACATGG GCCGGGAATT CGAGAATCTT TCATTTGGTG CATTGGCCGG GAATTCGAAA ATCTTCATT TGGTGCATTG GCCGGGAAAC AGCGCGACCA	540 600
40	CCCAGAGGTC CTAGACCCAC TTAGAGGTAA GATTCTTGT TCTGTTTGG TCTGATGTCT GTGTTCTGAT GTCTGTGTT TGTTCTAAG TCTGGTGCAG TCGCAGTTTC AGTTTGCGG	660 720
45	ACGCTCAGTG AGACCGCGCT CCGAGAGGGA GTGCGGGGTG GATAAGGATA GACGTGTCCA GGTGTCCACC GTCCGTTCGC CCTGGGAGAC GTCCCAGGAG GAACAGGGGA GGATCAGGGGA	780 840
50	CGCCTGGTGG ACCCCTTGA AGGCCAAGAG ACCATTTGGG GTTGCGAGAT CGTGGGTTCG AGTCCCACCT CGTGCCCAGT TGCGAGATCG TGGGTTCGAG TCCCACCTCG TGTTTGTTG	900 960
55	CGAGATCGTG GGTTCGAGTC CCACCTCGCG TCTGGTCACG GGATCGTGGG TTCGAGTCCC ACCTCGTGTGTT TGTTGCGAG ATCGTGGGTT CGAGTCCCAC CTCGCGTCTG GTCACGGGAT	1020 1080
60	CGTGGGTTCG AGTCCCACCT CGTGCAGAGG GTCTCAATTG GCCGGCCTTA GAGAGGCCAT CTGATTCTTC TGTTCTCT TTGTTCTTA GTCTCGTGTGTC CGCTCTTGTGTT GTGACTACTG	1140 1200
65	TTTTCTAAA AATGGGACAA TCTGTGTCCA CTCCCCTTTC TCTGACTCTG GTTCTGTGCG TTGGTAATT TGTTGTTTA CGTTGTTTT TGTGAGTCGT CTATGTTGTC TGTTACTATC	1260 1320
70	TTGTTTTGT TTGTTGTTA CGGTTCTGT GTGTGTCTTG TGTGTCTCTT TGTGTTCAGA CTTGGACTGA TGACTGACGA CTGTTTTAA GTTATGCCTT CTAAAATAAG CCTAAAAATC	1380 1440
75	CTGTCAGATC CCTATGCTGA CCACCTCCCT TCAGATCAAC AGCTGCCCTT ACTCGAGCTC AAGCTTCGAA TTCTGCAGTC GACGGTACCG CGGGGATCAA TTCCGCCCTT CCCCTAACGT	1500 1560

	TACTGGCCGA AGCCGCTTGG AATAAGGCCG GTGTGCCTT GTCTATATGT TATTTCCAC	1620
	CATATTGCCG TCTTTGGCA ATGTGAGGGC CCGGAAACCT GGCCCTGTCT TCTTGACGAG	1680
5	CATTCCTAGG GGTCTTCCC CTCTGCCAA AGGAATGCAA GGTCTGTTGA ATGTCGTGAA	1740
	GGAAGCAGTT CCTCTGGAAG CTTCTGAAG ACAAACACG TCTGTAGCGA CCCTTGAG	1800
10	GCAGCGGAAC CCCCCACCTG GCGACAGGTG CCTCTGCGC CAAAAGCCAC GTGTATAAGA	1860
	TACACCTGCA AAGGCGGCAC AACCCCAGTG CCACGTTGTG AGTTGGATAG TTGTGGAAAG	1920
	AGTCAAATGG CTCTCCTCAA GCGTATTCAA CAAGGGGCTG AAGGATGCC AGAAGGTACC	1980
15	CCATTGTATG GGATCTGATC TGGGGCCTCG GTGCACATGC TTTACATGTG TTTAGTCGAG	2040
	GTTAAAAAAC GTCTAGGCC CCGAACACAC GGGGACGTGG TTTCCCTTG AAAAACACGA	2100
20	GCGGGATCAA TTCCGCCCCC CCCCTAACGT TACTGGCCGA AGCCGCTTGG AATAAGGCCG	2160
	GTGTGCCTT GTCTATATGT TATTTCCAC CATATTGCCG TCTTTGGCA ATGTGAGGGC	2220
	CCGGAAACCT GCCCCTGTCT TCTTGACGAG CATTCTAGG GGTCTTCCC CTCTGCCAA	2280
25	AGGAATGCAA GGTCTGTTGA ATGTCGTGAA GGAAGCAGTT CCTCTGGAAG CTTCTGAAG	2340
	ACAAACAAACG TCTGTAGCGA CCCTTGAG GCAGCGGAAC CCCCCACCTG GCGACAGGTG	2400
30	CCTCTGCGC CAAAAGCCAC GTGTATAAGA TACACCTGCA AAGGCGGCAC AACCCCAGTG	2460
	CCACGTTGTG AGTTGGATAG TTGTGGAAAG AGTCAAATGG CTCTCCTCAA GCGTATTCAA	2520
	CAAGGGCTG AAGGATGCC AGAAGGTACC CCATTGTATG GGATCTGATC TGGGGCCTCG	2580
35	GTGCACATGC TTTACATGTG TTTAGTCGAG GTTAAAAAAA CGTCTAGGCC CCCCAGCCA	2640
	CGGGGACGTG GTTTCCCTT GAAAAACACG ATACGGATC CACCGGTCGC CACCATGGGT	2700
40	AAAGGAGAAG AACTTTCAC AGGAGTTGTC CCAATTCTTG TTGAATTAGA TGGTGATGTT	2760
	AATGGGCACA AATTTCTGT CAGTGGAGAG GGTGAAGGTG ATGCAACATA CGGAAAACATT	2820
	ACCCCTAAAT TTATTTGCAC TACTGGAAAA CTACCTGTT CATGGCCAAC ACTTGTCACT	2880
45	ACTTCACTT ATGGTGTCA ATGCTTTCA AGATACCCAG ATCATATGAA ACGGCATGAC	2940
	TTTTCAAGA GTGCCATGCC CGAAGGTTAT GTACAGGAAA GAACTATATT TTTCAAAGAT	3000
	GACGGAACT ACAAGACACG TGCTGAAGTC AAGTTGAAG GTGATACCCT TGTAAATAGA	3060
50	ATCGAGTTAA AAGGTATTGA TTTAAAGAA GATGGAAACA TTCTGGACA CAAATTGGAA	3120
	TACAACATATA ACTCACACAA TGTATACATC ATGGCAGACA AACAAAAGAA TGGAACCAAA	3180
55	GTAACTTCA AAATTAGACA CAACATTGAA GATGGAAGCG TTCAACTAGC AGACCATTAT	3240
	CAACAAAATA CTCCAATTGG CGATGGCCCT GTCCTTTAC CAGACAACCA TTACCTGTCC	3300
	ACACAATCTG CCCTTCGAA AGATCCAAC GAAAAGAGAG ACCACATGGT CCTTCTTGAG	3360
60	TTTGTAACAG CTGCTGGAT TACACATGGC ATGGATGAAC TATACAAGTC CGGATCTAGA	3420
	TAACGTATC GATGGATCCG AAGGCGGGGA CAGCAGTGCA GTGGTGGACA GAAAGCAAGT	3480
65	GATCTAGGCC AGCAGCCTCC CTAAAGGGAC TTCAGCCCAC AAAGCCAAAC TTGTGGCTT	3540
	AATACAAGCT CTGTAATGG TAAAAAAAAAA AAAGTCTACA CGGACAGCAG GTATGCTTT	3600

	GCCACTGTAC AGAGCAATAT ACAGACAAAG AGAACTGTTG ACATCTGCAG AGAAAGACCT	3660
5	AAGATGCTGT GGCTAAAAGA AATCAGATGG CAAATCTAAC CGCCCAGGCA TCCTAAAGAG	3720
	CAATGATCCT GACAGTCTGA AGACTATCAA GTTATAGACA AATTAAGACT GGTAAAAAAA	3780
	ACCCTGTATA AAATAGTAAA AACTGAAAAA AGAAAACTAG TCCTCTCATG AGAAGACAGA	3840
10	CCTGACATCT ACTGAAAAAT AGACTTTACT GGAAAAAATA TGTGTATGAA TACCTTCTAG	3900
	TTTTTGAA CGTTCTCAAG ATGGATAAAA GCTTTTCCTT GTAAAACGAG ACTGATCAGA	3960
15	TAGTCATCAA GAAGATTGTT AAAGAAAATT TTCCAAGGTT CGGAGTGCCA AAAGCAATAG	4020
	TGTCAGATAA TGGTCCTGCC TTTGTTGCC AGGTAAGTCA GGGTGTGCC AAGTATTAG	4080
	AGGTCAAATG AAAATTCCAT TGTGTGTACA GACCTCAGAG CTCAGGAAAG ATAAAAAAGA	4140
20	ATAAATAAAA CTCTAAACAG ACCTTGACAA AATTAATCCT AGAGACTGGC ACAGACTTAC	4200
	TTGGTACTCC TTCCCCTTGC CCTATTTAGA ACTGAGAATA CTCCCTCTTG ATTGGTTTT	4260
25	ACTCTTTTA AGATCCTTTA TGGGGCTCCT ATGCCATCAC TGTCTTAAAT GATGTGTTA	4320
	AACCTATGTT GTTATAATAA TGATCTATAT GTTAAGTTAA AAGGCTTGCA GGTGGTGCAG	4380
	AAAGAAGTCT GGTCACAACG GGCTACAGTG AACAAAGCTGG GTACCCCAAG GACATCTTAC	4440
30	CAGTTCCAGC CAGAGATCTG ATCTACGATC CCCGGTCA CCCGGTCA CCCTGTGGAA	4500
	TGTGTGTCAG TTAGGGTGTG GAAAGTCCCC AGGCTCCCCA GCAGGCAGAA GTATGCAAAG	4560
35	CATGCATCTC AATTAGTCAG CAACCAGGTG TGGAAAGTCC CCAGGCTCCC CAGCAGGCAG	4620
	AAGTATGCAA AGCATGCATC TCAATTAGTC AGCAACCATA GTCCCGCCCC TAACTCCGCC	4680
	CATCCCGCCC CTAACTCCGC CCAGTTCCGC CCATTCTCCG CCCCATGGCT GACTAATT	4740
40	TTTTATTTAT GCAGAGGCCG AGGCCGCCCTC GGCTCTGAG CTATTCCAGA AGTAGTGAGG	4800
	AGGCTTTTT GGAGGCCTAG GCTTTGCCAA AAAGCTTCAC GCTGCCGCAA GCACTCAGGG	4860
	CGCAAGGGCT GCTAAAGGAA GCGGAACACG TAGAAAGCCA GTCCGCAGAA ACGGTGCTGA	4920
45	CCCCGGATGA ATGTCAGCTA CTGGGCTATC TGGACAAGGG AAAACGCAAG CGCAAAGAGA	4980
	AAGCAGGTAG CTTGCAGTGG GCTTACATGG CGATAGCTAG ACTGGGCGGT TTTATGGACA	5040
	GCAAGCGAAC CGGAATTGCC AGCTGGGCG CCCTCTGGTA AGGTTGGAA GCCCTGCAA	5100
50	GTAAACTGGA TGGCTTCCTT GCCGCCAAGG ATCTGATGGC GCAGGGGATC AAGATCTGAT	5160
	CAAGAGACAG GATGAGGATC GTTTCGCATG ATTGAACAAG ATGGATTGCA CGCAGGTTCT	5220
55	CCGGCCGCTT GGGTGGAGAG GCTATTCCGC TATGACTGGG CACAACAGAC AATCGGCTGC	5280
	TCTGATGCCG CGTGTCCG GCTGTCAGCG CAGGGCGCC CGGTTCTTT TGTCAAGACC	5340
60	GACCTGTCCG GTGCCCTGAA TGAACATGCAG GACGAGGCAG CGCGGCTATC GTGGCTGGCC	5400
	ACGACGGGCG TTCCTTGCGC AGCTGTGCTC GACGTTGTCA CTGAAGCGGG AAGGGACTGG	5460
	CTGCTATTGG GCGAAGTGCC GGGGCAGGAT CTCCTGTCAT CTCACCTTGC TCCTGCCGAG	5520
65	AAAGTATCCA TCATGGCTGA TGCAATGCGG CGGCTGCATA CGCTTGATCC GGCTACCTGC	5580

CCATTCGACC ACCAAGCGAA ACATCGCATE GAGCGAGCAC GTACTCGGAT GGAAGCCGGT 5640  
 CTTGTCGATC AGGATGATCT GGACGAAGAG CATCAGGGC TCGCGCCAGC CGAACTGTT 5700  
 5 GCCAGGCTCA AGGCAGCGCAT GCCCGACGGC GAGGATCTCG TCGTGACCCA TGGCGATGCC 5760  
 TGCTTGCAGA ATATCATGGT GGAAAATGGC CGCTTTCTG GATTCACTGA CTGTGGCCGG 5820  
 10 CTGGGTGTGG CGGACCGCTA TCAGGACATA GCGTTGGCTA CCCGTGATAT TGCTGAAGAG 5880  
 CTTGGGGCG AATGGGCTGA CCGCTTCCTC GTGCTTTACG GTATCGCCGC TCCCGATTG 5940  
 CAGCGCATCG CCTTCTATCG CCTTCTTGAC GAGTTCTTCT GAGCAGGGACT CTGGGGTTG 6000  
 15 AAATGACCGA CCAAGCGACG CCCAACCTGC CATCACGAGA TTTCGATTCC ACCGCCGCCT 6060  
 TCTATGAAAG GTTGGGCTTC GGAATCGTT TCCGGGACGG AATTGTAAT CTGCTGCTTG 6120  
 20 CAAACAAAAA AACCAACCGCT ACCAGCGGTG GTTGTTGCG CGGATCAAGA GCTACCAACT 6180  
 CTTTTCCGA AGGTAACTGG CTTCAGCAGA GCGCAGATAAC CAAATACTGT CCTTCTAGTG 6240  
 TAGCCGTAGT TAGGCCACCA CTTCAAGAAC TCTGTAGCAC CGCCTACATA CCTCGCTCTG 6300  
 25 CTAATCCTGT TACCAAGTGGC TGCTGCCAGT GGCATAAGT CGTGTCTTAC CGGGTTGGAC 6360  
 TCAAGACGAT AGTTACCGGA TAAGGGCGAG CGGTGGGCT GAACGGGGGG TTCTGTGCACA 6420  
 30 CAGCCCAGCT TGGAGCGAAC GACCTACACC GAACTGAGAT ACCTACAGCG TGAGCATTGA 6480  
 GAAAGCGCCA CGCTTCCCGA AGGGAGAAAG GCGGACAGGT ATCCGGTAAG CGGCAGGGTC 6540  
 GGAACAGGAG AGCGCACGAG GGAGCTTCCA GGGGAAACG CCTGGTATCT TTATAGTCCT 6600  
 35 GTCGGGTTTC GCCACCTCTG ACTTGAGCGT CGATTTTGT GATGCTCGTC AGGGGGCGG 6660  
 AGCCTATGGA AAAACGCCAG CAACGCCAG ATGCGCCGCC TCGAGTACAC CTGCGTCATG 6720  
 40 CTGAGACCCCT CAAGCCTCAC TAAAAGGGTC CCTGCCTAGT TCTGTTACT AATCTGCCTT 6780  
 ATTCTGTTTT TGTTCCCATG TTAAAGATAG AGTAAATGCA GTATTCTCCA CATAGAGATA 6840  
 TAGACTTCTG AAATTCTAAG ATTAGAATTA TTTACAAGAA GAAGTGGGAA A 6891

## 45 (2) INFORMATION FOR SEQ ID NO:17:

- (i) SEQUENCE CHARACTERISTICS:  
 (A) LENGTH: 6321 base pairs  
 (B) TYPE: nucleic acid  
 (C) STRANDEDNESS: single  
 (D) TOPOLOGY: linear

50 (ii) MOLECULE TYPE: DNA (genomic)

55

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:17:

TGAAGAATAA AAAATTACTG GCCTCTTGTG AGAACATGAA CTTTCACCTC GGAGCCACC 60  
 60 CCCTCCCCATC TGGAAAACAT ACTTGAGAAA AACATTTCT GGAACAAACCA CAGAATGTT 120  
 CAACAGGCCA GATGTATTGC CAAACACAGG ATATGACTCT TTGGTTGAGT AAATTGTGG 180  
 65 TTGTTAAACT TCCCCTATTG CCTCCCCATT CCCCTCCCA GTTGTGGTT TTTCCCTTA 240

	AAAGCTTGTG AAAAATTGA GTCGTCGTCG AGACTCCTCT ACCCTGTGCA AAGGTGTATG	300
	AGTTTCGACC CCAGAGCTCT GTGTGCTTTC TGTTGCTGCT TTATTCGAC CCCAGAGCTC	360
5	TGGTCTGTGT GCTTCATGT CGCTGCTTTA TTAAATCTTA CCTTCTACAT TTTATGTATG	420
	GTCTCAGTGT CTTCTGGGT ACGCGGCTGT CCCGGGACTT GAGTGTCTGA GTGAGGGTCT	480
10	TCCCTCGAGG GTCTTCATT TGGTACATGG GCCGGGAATT CGAGAATCTT TCATTTGGTG	540
	CATTGGCCGG GAATTCGAAA ATCTTCATT TGGTGCATTG GCCGGGAAAC AGCGCGACCA	600
	CCCAGAGGTC CTAGACCCAC TTAGAGGTAA GATTCTTGT TCTGTTTGG TCTGATGTCT	660
15	GTGTTCTGAT GTCTGTGTTG TGTTTCTAAG TCTGGTGCAG TCGCAGTTTC AGTTTGCGG	720
	ACGCTCAGTG AGACCGCGCT CCGAGAGGGA GTGCCGGGTG GATAAGGATA GACGTGTCCA	780
20	GGTGTCCACC GTCCGTTCGC CCTGGGAGAC GTCCCAGGAG GAACAGGGGA GGATCAGGGA	840
	CGCCTGGTGG ACCCCTTGA AGGCCAAGAG ACCATTGGG GTTGCAGAT CGTGGGTTCG	900
	AGTCCCACCT CGTGCCCAGT TCGAGATCG TGGGTTCGAG TCCCACCTCG TGTTTGTTG	960
25	CGAGATCGTG GGTCGAGTC CCACCTCGCG TCTGGTCACG GGATCGTGGG TTCGAGTCCC	1020
	ACCTCGTGT TTGTTGCGAG ATCGTGGTT CGAGTCCCAC CTCGCGTCTG GTCACGGGAT	1080
30	CGTGGGTTCG AGTCCCACCT CGTGCAGAGG GTCTCAATTG GCCGGCCTTA GAGAGGCCAT	1140
	CTGATTCTTC TGTTTCTCT TTTGTCTTA GTCTCGTGT CGCTCTTGTG TGACTACTG	1200
	TTTTCTAAA AATGGGACAA TCTGTGTCCA CTCCCTTTC TCTGACTCTG GTTCTGTGCG	1260
35	TTGGTAATT TGTTGTTTA CGTTGTTTT TGTGAGTCGT CTATGTTGTC TGTTACTATC	1320
	TTGTTTTGT TTGTGGTTA CGGTTCTGT GTGTGTCTTG TGTGTCTCTT TGTGTTCAGA	1380
	CTTGGACTGA TGACTGACGA CTGTTTTAA GTTATGCCTT CTAAAATAAG CCTAAAAATC	1440
40	CTGTCAGATC CCTATGCTGA CCACCTCCTT TCAGATCAAC AGCTGCCCTT ACTCGAGCTC	1500
	AAGCTTCGAA TTCTGCAGTC GACGGTACCG CGGGGATCAA TTCCGCCCCC CCCCTAACGT	1560
45	TACTGCCGA AGCCGTTGG AATAAGGCCG GTGTGCGTT GTCTATATGT TATTTCCAC	1620
	CATATTGCCG TCTTTGGCA ATGTGAGGGC CCGGAAACCT GGCCCTGTCT TCTTGACGAG	1680
50	CATTCCTAGG GGTCTTCCTT CTCTGCCAA AGGAATGCAA GGTCTGTTGA ATGTCGTGAA	1740
	GGAAGCAGTT CCTCTGGAAG CTTCTGAAG ACAAAACAACG TCTGTAGCGA CCCTTGCAG	1800
	GCAGCGGAAC CCCCCACCTG GCGACAGGTG CCTCTGCCGC CAAAAGCCAC GTGTATAAGA	1860
55	TACACCTGCA AAGGGGCAC AACCCCAGTG CCACGTTGT AGTTGGATAG TTGTGGAAAG	1920
	AGTCAAATGG CTCTCCTCAA GCGTATTCAA CAAGGGCTG AAGGATGCC AGAAGGTACC	1980
60	CCATTGTATG GGATCTGATC TGGGGCCTCG GTGCACATGC TTTACATGTG TTTAGTCGAG	2040
	GTTAAAAAAA CGTCTAGGCC CCCCCAACCA CGGGGACGTG GTTTCTTT GAAAAACACG	2100
	ATACGGGATC CACCGGTGCG CACCATGGGT AAAGGAGAAG AACTTTCAC AGGAGTTGTC	2160
65	CCAATTCTTG TTGAATTAGA TGGTGTGTT AATGGGCACA AATTTCTGT CAGTGGAGAG	2220

	GGTGAAGGTG ATGCAACATA CGGAAAAC TTATTTGCAC TACTGGAAAA	2280
	CTACCTGTT CATGCCAAC ACTTGTCACT ACTTTCAC TTGGTGTCA ATGCTTTCA	2340
5	AGATACCCAG ATCATATGAA ACGGCATGAC TTTTCAAGA GTGCCATGCC CGAAGGTTAT	2400
	GTACAGGAAA GAACTATATT TTTCAAAGAT GACGGGAAC ACAAGACACG TGCTGAAGTC	2460
10	AAGTTTGAAG GTGATACCCCT TGTTAATAGA ATCGAGTTAA AAGGTATTGA TTTTAAAGAA	2520
	GATGGAAACA TTCTTGACA CAAATTGGAA TACAACATA ACTCACACAA TGTATACATC	2580
	ATGGCAGACA AACAAAAGAA TGGAACCAAA GTTAACCTCA AAATTAGACA CAACATTGAA	2640
15	GATGGAAGCG TTCAACTAGC AGACCATTAT CAACAAAATA CTCCAATTGG CGATGCCCT	2700
	GTCCTTTAC CAGACAACCA TTACCTGTCC ACACAATCTG CCCTTCGAA AGATCCCAAC	2760
20	GAAAAGAGAG ACCACATGGT CCTTCTTGAG TTTGTAACAG CTGCTGGGAT TACACATGGC	2820
	ATGGATGAAC TATACAAGTC CGGATCTAGA TAACTGTATC GATGGATCCG AAGGCGGGGA	2880
	CAGCAGTGCA GTGGTGGACA GAAAGCAAGT GATCTAGGCC AGCAGCCTCC CTAAAGGGAC	2940
25	TTCAGCCCAC AAAGCCAAAC TTGTGGCTTT AATACAAGCT CTGTAATGG TAAAAAAA	3000
	AAAGTCTACA CGGACAGCAG GTATGCTTT GCCACTGTAC AGAGCAATAT ACAGACAAAG	3060
30	AGAACTGTTG ACATCTGCAG AGAAAGACCT AAGATGCTGT GGCTAAAAGA AATCAGATGG	3120
	CAAATCTAAC CGCCCGAGGCA TCCTAAAGAG CAATGATCCT GACAGTCTGA AGACTATCAA	3180
	GTTATAGACA AATTAAGACT GGTAAAAAAA ACCCTGTATA AAATAGTAAA AACTGAAAAA	3240
35	AGAAAAACTAG TCCTCTCATG AGAAGACAGA CCTGACATCT ACTGAAAAAT AGACTTTACT	3300
	GGAAAAAAATA TGTGTATGAA TACCTCTAG TTTTGTGAA CGTTCTCAAG ATGGATAAAA	3360
40	GCTTTCTT GTAAAACGAG ACTGATCAGA TAGTCATCAA GAAGATTGTT AAAGAAAATT	3420
	TTCCAAGGTT CGGAGTGCCA AAAGCAATAG TGTCAGATAA TGGTCCTGCC TTTGTTGCC	3480
	AGGTAAGTCA GGGTGTGGCC AAGTATTTAG AGGTCAAATG AAAATTCCAT TGTGTGTACA	3540
45	GACCTCAGAG CTCAGGAAAG ATAAAAAAAGA ATAAATAAAA CTCTAACAG ACCTTGACAA	3600
	AATTAATCCT AGAGACTGGC ACAGACTTAC TTGGTACTCC TTCCCCCTGC CCTATTTAGA	3660
	ACTGAGAATA CTCCCTCTTG ATTGGTTTT ACTCTTTTA AGATCCTTA TGGGGCTCCT	3720
50	ATGCCATCAC TGTCTTAAAT GATGTGTTA AACCTATGTT GTTATAATAA TGATCTATAT	3780
	GTTAAGTTAA AAGGCTTGCA GGTGGTGCAG AAAGAAGTCT GGTACAAACT GGCTACAGTG	3840
55	AAACAAGCTGG GTACCCCAAG GACATCTTAC CAGTTCCAGC CAGAGATCTG ATCTACGATC	3900
	CCCAGGGTCGA CCCGGGTCGA CCCTGTGGAA TGTGTGTCAAG TTAGGGTGTG GAAAGTCCCC	3960
	AGGCTCCCCA GCAGGCAGAA GTATGCAAAG CATGCATCTC AATTAGTCAG CAACCAGGTG	4020
60	TGGAAAGTCC CCAGGCTCCC CAGCAGGCAG AAGTATGCAA AGCATGCATC TCAATTAGTC	4080
	AGCAACCATA GTCCCCCCCC TAACTCCGCC CATCCCCCCC CTAACTCCGC CCAGTTCCGC	4140
65	CCATTCTCCG CCCCATGGCT GACTAATTAA TTTTATTAT GCAGAGGCCG AGGCCGCCTC	4200
	GGCCTCTGAG CTATTCCAGA AGTAGTGTGAGG AGGCTTTTT GGAGGCCCTAG GCTTTGCAA	4260

	AAAGCTTCAC GCTGCCGCAA GCACTCAGGG CGCAAGGGCT GCTAAAGGAA GCGGAACACG	4320
5	TAGAAAGCCA GTCCGCAGAA ACGGTGCTGA CCCC GGATGA ATGTCAGCTA CTGGGCTATC	4380
	TGGACAAGGG AAAACGCAAG CGCAAAGAGA AAGCAGGTAG CTTGCAGTGG GCTTACATGG	4440
	CGATAGCTAG ACTGGGCGGT TTTATGGACA GCAAGCGAAC CGGAATTGCC AGCTGGGCG	4500
10	CCCTCTGGTA AGGTTGGAA GCCCTGCAAA GTAAACTGGA TGGCTTCCTT GCCGCCAAGG	4560
	ATCTGATGGC GCAGGGGATC AAGATCTGAT CAAGAGACAG GATGAGGATC GTTTCGCATG	4620
15	ATTGAACAAG ATGGATTGCA CGCAGGTTCT CCGGCCGCTT GGGTGGAGAG GCTATTGGC	4680
	TATGACTGGG CACAACAGAC AATCGGCTGC TCTGATGCCG CCGTGTTCGG GCTGTCAGCG	4740
	CAGGGGCGCC CGGTTCTTT TGTCAAGACC GACCTGTCCG GTGCCCTGAA TGAAC TGCACTGCAG	4800
20	GACGAGGCAG CGCGGCTATC GTGGCTGGCC ACGACGGGCG TTCCTTGCGC AGCTGTGCTC	4860
	GACGTTGTCA CTGAAGCGGG AAGGGACTGG CTGCTATTGG GCGAAGTGCC GGGGCAGGAT	4920
25	CTCCTGTCAT CTCACCTTGC TCCTGCCGAG AAAGTATCCA TCATGGCTGA TGCAATGCGG	4980
	CGGCTGCATA CGCTTGATCC GGCTACCTGC CCATTCGACC ACCAAGCGAA ACATCGCATC	5040
	GAGCGAGCAC GTACTCGGAT GGAAGCCGGT CTTGTCGATC AGGATGATCT GGACGAAGAG	5100
30	CATCAGGGGC TCGCGCCAGC CGAACTGTTC GCCAGGCTCA AGGCGCGCAT GCCCGACGGC	5160
	GAGGATCTCG TCGTGACCCA TGGCGATGCC TGCTGCCGA ATATCATGGT GGAAAATGGC	5220
35	CGCTTTCTG GATTCATCGA CTGTGGCCGG CTGGGTGTGG CGGACCGCTA TCAGGACATA	5280
	GCGTTGGCTA CCCGTGATAT TGCTGAAGAG CTTGGCGCG AATGGGCTGA CCGCTTCCTC	5340
	GTGCTTTACG GTATCGCCGC TCCCGATTG CAGCGCATCG CCTTCTATCG CCTTCTTGAC	5400
40	GAGTTCTTCT GAGCGGGACT CTGGGGTTCG AAATGACCGA CCAAGCGACG CCCAACCTGC	5460
	CATCACGAGA TTTCGATTCC ACCGCCGCCT TCTATGAAAG GTGGGCTTC GGAATCGTTT	5520
	TCCGGGACGG ATTCTGTAAT CTGCTGCTTG CAAACAAAA AACCAACCGCT ACCAGCGGTG	5580
45	GTTTGTTGC CGGATCAAGA GCTACCAA CTTTTCCGA AGGTAAC TGG CTT CAGCAGA	5640
	GCGCAGATA CAAATACTGT CCTTCTAGTG TAGCCGTAGT TAGGCCACCA CTTCAAGAAC	5700
50	TCTGTAGCAC CGCCTACATA CCTCGCTCTG CTAATCCTGT TACCA GTGGC TGCTGCCAGT	5760
	GGCGATAAGT CGTGTCTTAC CGGGTTGGAC TCAAGACGAT AGTTACCGGA TAAGGCGCAG	5820
	CGGTGGGCT GAACGGGGGG TTCGTGCACA CAGCCCAGCT TGGAGCGAAC GACCTACACC	5880
55	GAAC TGAGAT ACCTACAGCG TGAGCATTGA GAAAGCGCCA CGCTTCCCGA AGGGAGAAAG	5940
	CGGGACAGGT ATCCGGTAAG CGGCAGGGTC GGAACAGGAG AGCGCACGAG GGAGCTTCCA	6000
60	GGGGGAAACG CCTGGTATCT TTATAGTCCT GTGGGTTTC GCCACCTCTG ACTTGAGCGT	6060
	CGATTGGT GATGCTCGTC AGGGGGCGG AGCCTATGGA AAAACGCCAG CAACGCCAG	6120
	ATGCGCCGCC TCGAGTACAC CTGCGTCATG CTGAGACCCCT CAAGCCTCAC TAAAAGGGTC	6180
65	CCTGCCTAGT TCTGTTTACT AATCTGCCTT ATTCTGTTT TGTTCCCATG TTAAAGATAG	6240

AGTAAATGCA GTATTCTCCA CATAGAGATA TAGACTTCGT AAATTCTAAG ATTAGAATTA	6300
TTTACAAGAA GAAGTGGGGA A	6321

## 5 (2) INFORMATION FOR SEQ ID NO:18:

- (i) SEQUENCE CHARACTERISTICS:
  - (A) LENGTH: 5754 base pairs
  - (B) TYPE: nucleic acid
  - (C) STRANDEDNESS: single
  - (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: DNA (genomic)

15

## (xi) SEQUENCE DESCRIPTION: SEQ ID NO:18:

20 TGAAGAATAA AAAATTACTG GCCTCTTGTG AGAACATGAA CTTTCACCTC GGAGCCCACC	60
CCCTCCCCATC TGGAAAACAT ACTTGAGAAA AACATTTCT GGAACAAACCA CAGAATGTTT	120
25 CAACAGGCCA GATGTATTGC CAAACACAGG ATATGACTCT TTGGTTGAGT AAATTGTGG	180
TTGTTAAACT TCCCCTATTG CCTCCCCATT CCCCTCCCA GTTGTGGTT TTTCTTTA	240
AAAGCTTGTG AAAAATTGAGA GTCGTCGTCG AGACTCCTCT ACCCTGTGCA AAGGTGTATG	300
30 AGTTTCGACC CCAGAGCTCT GTGTGCTTTC TGTTGCTGCT TTATTCGAC CCCAGAGCTC	360
TGGTCTGTGT GCTTCATGT CGCTGCTTTA TTAAATCTTA CCTTCTACAT TTTATGTATG	420
35 GTCTCAGTGT CTTCTTGGGT ACGCGGCTGT CCCGGGACTT GAGTGTCTGA GTGAGGGTCT	480
TCCCTCGAGG GTCTTCATT TGGTACATGG GCCGGGAATT CGAGAATCTT TCATTTGGTG	540
CATTGGCCGG GAATTTCGAAA ATCTTCATT TGGTGCATTG GCCGGGAAAC AGCGCGACCA	600
40 CCCAGAGGTC CTAGACCCAC TTAGAGGTAA GATTCTTGT TCTGTTTGG TCTGATGTCT	660
GTGTTCTGAT GTCTGTGTT TGTTCTAAG TCTGGTGCAGA TCGCAGTTTC AGTTTGCGG	720
45 ACGCTCAGTGT AGACCGCGCT CCGAGAGGGA GTGCGGGGTG GATAAGGATA GACGTGTCCA	780
GGTGTCCACC GTCCGTTCGC CCTGGGAGAC GTCCCAGGAG GAACAGGGGA GGATCAGGGA	840
CGCCTGGTGG ACCCCTTTGA AGGCCAAGAG ACCATTTGGG GTTGCAGAT CGTGGGTTCG	900
50 AGTCCCACCT CGTGCCCAGT TGCGAGATCG TGGGTTCGAG TCCCACCTCG TGTTTGTG	960
CGAGATCGTG GGTTCGAGTC CCACCTCGCG TCTGGTCACG GGATCGTGGG TTCGAGTCCC	1020
55 ACCTCGTGT TTGTTGCGAG ATCGTGGGTT CGAGTCCCAC CTCGCGTCTG GTCACGGGAT	1080
CGTGGGTTCG AGTCCCACCT CGTGCAGAGG GTCTCAATTG GCCGGCCTTA GAGAGGCCAT	1140
CTGATTCTTC TGTTTCTCT TTGTTCTTA GTCTCGTGT CGCTCTTGT GTGACTACTG	1200
60 TTGTTCTAAA AATGGGACAA TCTGTGTCCA CTCCCCTTTC TCTGACTCTG GTTCTGTCGC	1260
TTGGTAATTG TGTTTGTGTTA CGTTTGTGTT TGAGTGTGCT CTATGTTGTC TGTTACTATC	1320
TTGTTTTGT TTGTTGTTA CGGTTCTGT GTGTGTCTTG TGTGTCTCTT TGTGTTCAGA	1380
65 CTTGGACTGA TGACTGACGA CTGTTTTAA GTTATGCCTT CTAAAATAAG CCTAAAAATC	1440

	CTGTCAGATC CCTATGCTGA CCACCTCCTT TCAGATCAAC AGCTGCCCTT ACTCGAGCTC	1500
5	AAGCTTCGAA TTCTGCAGTC GACGGTACCG CGGGCCCGGG ATCCACCGGT CGCCACCATG	1560
	GGTAAAGGAG AAGAACTTTT CACAGGAGTT GTCCCAATT TC TTGTTGAATT AGATGGTGAT	1620
	GTAAATGGC ACAAAATTTT TGTCAGTGG AAGGGTGAAG GTGATGCAAC ATACGGAAAA	1680
10	CTTACCCCTTA AATTATTTG CACTACTGG AAAACTACCTG TTCCATGCC AACACTTGT	1740
	ACTACTTCA CTTATGGTGT TCAATGCTTT TCAAGATACC CAGATCATAT GAAACGGCAT	1800
15	GACTTTTCA AGAGTGCCAT GCCCGAAGGT TATGTACAGG AAAGAACTAT ATTTTCAA	1860
	GATGACGGGA ACTACAAGAC ACGTGCTGAA GTCAAGTTG AAGGTGATAC CCTTGTAA	1920
	AGAATCGAGT TAAAAGGTAT TGATTTAAA GAAGATGGAA ACATCTTGG ACACAAATTG	1980
20	GAATACAAC TAAACTCACA CAATGTATAAC ATCATGGCAG ACAAAACAAA GAATGGAACC	2040
	AAAGTTAACT TCAAAATTAG ACACAACATT GAAGATGGAA GCGTTCAACT AGCAGACC	2100
25	TATCAACAAA ATACTCCAAT TGGCGATGGC CCTGTCCTT TACCAGACAA CCATTACCTG	2160
	TCCACACAAT CTGCCCTTTC GAAAGATCCC AACGAAAAGA GAGACCACAT GGTCTTCTT	2220
	GAGTTTGTAA CAGCTGCTGG GATTACACAT GGCATGGATG AACTATACAA GTCCGGATCT	2280
30	AGATAACTGT ATCGATGGAT CCGAAGGCAGG GGACAGCAGT GCAGTGGTGG ACAGAAAGCA	2340
	AGTGATCTAG GCCAGCAGCC TCCCTAAAGG GACTTCAGCC CACAAAGCCA AACTTGTGGC	2400
35	TTTAATACAA GCTCTGTAAA TGGTAAAAAA AAAAAGTCT ACACGGACAG CAGGTATGCT	2460
	CTTGCCACTG TACAGAGCAA TATACAGACA AAGAGAACTG TTGACATCTG CAGAGAAAGA	2520
	CCTAAGATGC TGTGGCTAAA AGAAATCAGA TGGCAAATCT AACCGCCCAG GCATCCTAAA	2580
40	GAGCAATGAT CCTGACAGTC TGAAGACTAT CAAAGTTATAG ACAAAATTAAG ACTGGTAAA	2640
	AAAACCTGT ATAAAATAGT AAAAAGTCAA AAAAGAAAAC TAGTCCTCTC ATGAGAAAGAC	2700
	AGACCTGACA TCTACTGAAA AATAGACTTT ACTGGAAAAA ATATGTGTAT GAATACCTTC	2760
45	TAGTTTTGT GAACGTTCTC AAGATGGATA AAAGCTTTC CTTGTAAAAC GAGACTGATC	2820
	AGATAGTCAT CAAGAAGATT GTTAAAGAAA ATTTTCCAAG GTTCGGAGTG CCAAAAGCAA	2880
50	TAGTGTCAAGA TAATGGTCCT GCCTTGTG CCCAGGTAAG TCAGGGTGTG GCCAAGTATT	2940
	TAGAGGTCAA ATGAAAATTCA CATTGTGTGT ACAGACCTCA GAGCTCAGGA AAGATAAAA	3000
	AGAATAAAATA AACTCTAAA CAGACCTTGA CAAAATTAAT CCTAGAGACT GGCACAGACT	3060
55	TACTTGGTAC TCCTTCCCCCT TGCCCTATTT AGAACTGAGA ATACTCCCTC TTGATTGGT	3120
	TTTACTCTTT TTAAGATCCT TTATGGGCT CCTATGCCAT CACTGTCTTA AATGATGTGT	3180
60	TTAACCTAT GTTGTATCAA TAATGATCTA TATGTTAAGT TAAAAGGCTT GCAGGTGGT	3240
	CAGAAAGAAG TCTGGTCACA ACTGGCTACA GTGAACAAAGC TGGGTACCCC AAGGACATCT	3300
	TACCAGTTCC AGCCAGAGAT CTGATCTACG ATCCCCGGGT CGACCCGGGT CGACCCGTG	3360
65	GAATGTGTGT CAGTTAGGGT GTGGAAAGTC CCCAGGCTCC CCAGCAGGCA GAAGTATGCA	3420

	AAGCATGCAT	CTCAATTAGT	CAGCAACCAG	GTGTGGAAAG	TCCCCAGGCT	CCCCAGCAGG	3480
	CAGAAGTATG	CAAAGCATGC	ATCTCAATTA	GTCAGCAACC	ATAGTCCCGC	CCCTAACTCC	3540
5	GCCCATCCCG	CCCCTAACTC	CGCCCAGTTC	CGCCCCATTCT	CCGCCCATG	GCTGACTAAT	3600
	TTTTTTTATT	TATGCAGAGG	CCGAGGCCGC	CTCGGCCTCT	GAGCTATTCC	AGAAGTAGTG	3660
10	AGGAGGCTT	TTTGGAGGCC	TAGGCTTTG	CAAAAAGCTT	CACGCTGCCG	CAAGCACTCA	3720
	GGGCGCAAGG	GCTGCTAAAG	GAAGCGGAAC	ACGTAGAAAG	CCAGTCCGCA	GAAACGGTGC	3780
	TGACCCCGGA	TGAATGTCAG	CTACTGGGCT	ATCTGGACAA	GGGAAAACGC	AAGCGCAAAG	3840
15	AGAAAGCAGG	TAGCTTGCAG	TGGGCTTACA	TGGCGATAGC	TAGACTGGGC	GGTTTTATGG	3900
	ACAGCAAGCG	AACCGGAATT	GCCAGCTGGG	GCGCCCTCTG	GTAAGGTTGG	GAAGCCCTGC	3960
20	AAAGTAAACT	GGATGGCTT	CTTGCCGCCA	AGGATCTGAT	GGCGCAGGGG	ATCAAGATCT	4020
	GATCAAGAGA	CAGGATGAGG	ATCGTTTCGC	ATGATTGAAC	AAGATGGATT	GCACGCAGGT	4080
	TCTCCGGCCG	CTTGGGTGGA	GAGGCTATTTC	GGCTATGACT	GGGCACAACA	GACAATCGGC	4140
25	TGCTCTGATG	CCGCCGTGTT	CCGGCTGTCA	GCGCAGGGGC	GCCC GGTTCT	TTTTGTCAAG	4200
	ACCGACCTGT	CCGGTGCCTCT	GAATGAACTG	CAGGACGAGG	CAGCGCGGCT	ATCGTGCTG	4260
30	GCCACGACGG	GCGTTCCCTG	CGCAGCTGTG	CTCGACGTTG	TCACTGAAGC	GGGAAGGGAC	4320
	TGGCTGCTAT	TGGGCGAAGT	GCCGGGGCAG	GATCTCCTGT	CATCTCACCT	TGCTCCTGCC	4380
	GAGAAAGTAT	CCATCATGGC	TGATGCAATG	CGGC GGCTGC	ATACGCTTGA	TCCGGCTACC	4440
35	TGCCCATTCG	ACCACCAAGC	GAAACATCGC	ATCGAGCGAG	CACGTACTCG	GATGGAAGCC	4500
	GGTCTTGTG	ATCAGGATGA	TCTGGACGAA	GAGCATCAGG	GGCTCGCGCC	AGCCGAAC TG	4560
40	TTCGCCAGGC	TCAAGGC GCG	CATGCCCGAC	GGCGAGGATC	TCGTCGTGAC	CCATGGCGAT	4620
	GCCTGCTTGC	CGAATATCAT	GGTGAAAAT	GGCGCCTTT	CTG GATT CAT	CGACTGTGGC	4680
	CGGCTGGGTG	TGGCGGACCG	CTATCAGGAC	ATAGCGTTGG	CTACCCGTGA	TATTGCTGAA	4740
45	GAGCTTGGCG	GCGAATGGGC	TGACCGCTTC	CTCGTGCCTT	ACGGTATCGC	CGCTCCCGAT	4800
	TCGCAGCGCA	TCGCCTTCTA	TCGCCTTCTT	GACGAGTTCT	TCTGAGCGGG	ACTCTGGGT	4860
	TCGAAATGAC	CGACCAAGCG	ACGCCAAC	TGCCATCACG	AGATTCGAT	TCCACCGCCG	4920
50	CCTTCTATGA	AAGGTTGGGC	TTCGGAATCG	TTTCCGGGA	CGGAATT CGT	AATCTGCTGC	4980
	TTGCAAACAA	AAAAACCACC	GCTACCAGCG	GTGGTTGTT	TGCCGGATCA	AGAGCTACCA	5040
	ACTCTTTTC	CGAAGGTAAC	TGGCTTCAGC	AGAGCGCAGA	TACCAAATAC	TGCTCCTCTA	5100
55	GTGTAGCCGT	AGTTAGGCCA	CCACTTCAAG	AACTCTGTAG	CACCGCCTAC	ATACCTCGCT	5160
	CTGCTAATCC	TGTTACCAGT	GGCTGCTGCC	AGTGGCGATA	AGTCGTGTCT	TACCGGGTTG	5220
60	GACTCAAGAC	GATAGTTACC	GGATAAGCG	CAGCGGTCGG	GCTGAACGGG	GGGTTCGTGC	5280
	ACACAGCCCA	GCTTGGAGCG	AACGACCTAC	ACCGA ACTGA	GATA CCTACA	GGGTGAGCAT	5340
65	TGAGAAAGCG	CCACGCTTCC	CGAAGGGAGA	AAGGC GGACA	GGTATCCGGT	AAGCGGCAGG	5400
	GTCGGAACAG	GAGAGCGCAC	GAGGGAGCTT	CCAGGGGGAA	ACGCCTGGTA	TCTTTATAGT	5460

CCTGTCGGGT TTCGCCACCT CTGACTTGAG CGTCGATTT TGTGATGCTC GTCAGGGGG 5520  
 CGGAGCCTAT GGAAAAACGC CAGCAACGCC GAGATGCGCC GCCTCGAGTA CACCTGCGTC 5580  
 5 ATGCTGAGAC CCTCAAGCCT CACTAAAAGG GTCCCTGCCT AGTTCTGTT ACTAATCTGC 5640  
 CTTATTCTGT TTTGTTCCC ATGTTAAAGA TAGAGTAAAT GCAGTATTCT CCACATAGAG 5700  
 10 ATATAGACTT CTGAAATTCT AAGATTAGAA TTATTTACAA GAAGAAGTGG GGAA 5754

## (2) INFORMATION FOR SEQ ID NO:19:

- 15 (i) SEQUENCE CHARACTERISTICS:  
 (A) LENGTH: 5754 base pairs  
 (B) TYPE: nucleic acid  
 (C) STRANDEDNESS: single  
 (D) TOPOLOGY: linear

20 (ii) MOLECULE TYPE: DNA (genomic)

## 25 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:19:

TGAAGAATAA AAAATTACTG GCCTCTTGTG AGAACATGAA CTTTCACCTC GGAGCCCACC 60  
 CCCTCCCATC TGGAAAACAT ACTTGAGAAA AACATTTCT GGAACAACCA CAGAATGTTT 120  
 30 CAACAGGCCA GATGTATTGC CAAACACAGG ATATGACTCT TTGGTTGAGT AAATTGTGG 180  
 TTGTTAAACT TCCCCTATTC CCTCCCCATT CCCCTCCCA GTTTGTGGTT TTTCTTTA 240  
 35 AAAGCTTGTG AAAAATTGGA GTCGTCGTCG AGACTCCTCT ACCCTGTGCA AAGGTGTATG 300  
 AGTTTCGACC CCAGAGCTCT GTGTGCTTTC TGTTGCTGCT TTATTCGAC CCCAGAGCTC 360  
 40 TGGTCTGTGT GCTTCATGT CGCTGCTTTA TTAAATCTTA CCTTCTACAT TTTATGTATG 420  
 GTCTCAGTGT CTTCTGGGT ACGCGGCTGT CCCGGGACTT GAGTGTCTGA GTGAGGGTCT 480  
 TCCCTCGAGG GTCTTCATT TGGTACATGG GCCGGGAATT CGAGAATCTT TCATTTGGTG 540  
 45 CATTGGCCGG GAATTGAAA ATCTTCATT TGGTGCATTG GCCGGGAAAC AGCGCGACCA 600  
 CCCAGAGGTC CTAGACCCAC TTAGAGGTAA GATTCTTGT TCTGTTTGG TCTGATGTCT 660  
 50 GTGTTCTGAT GTCTGTGTTC TGTTCTAAG TCTGGTGCAG TCGCAGTTTC AGTTTGCGG 720  
 ACGCTCAGTG AGACCGCGCT CCGAGAGGGA GTGCGGGGTG GATAAGGATA GACGTGTCCA 780  
 GGTGTCCACC GTCCGTTCGC CCTGGGAGAC GTCCCAGGAG GAACAGGGGA GGATCAGGG 840  
 55 CGCCTGGTGG ACCCCTTGA AGGCCAAGAG ACCATTTGGG GTTGCGAGAT CGTGGGTTCG 900  
 AGTCCCACCT CGTGCCCAGT TGCGAGATCG TGGGTTCGAG TCCCACCTCG TGTTTGTTG 960  
 60 CGAGATCGTG GGTTCGAGTC CCACCTCGCG TCTGGTCACG GGATCGTGGG TTGAGTCCC 1020  
 ACCTCGTGTGTT TTGTTGCGAG ATCGTGGGTT CGAGTCCCAC CTCCGCGTCTG GTCACGGGAT 1080  
 CGTGGGTTCG AGTCCCACCT CGTGCAGAGG GTCTCAATTG GCCGGCCTTA GAGAGGCCAT 1140  
 65 CTGATTCTTC TGGTTCTCT TTTGCTTTA GTCTCGTGTGTC CGCTCTTGTGTT GTGACTACTG 1200

	TTTTTCTAAA AATGGGACAA TCTGTGTCCA CTCCCCTTTC TCTGACTCTG GTTCTGTCGC	1260
	TTGGTAATTT TGTTTGTAA CGTTTGTAA TGTTGAGTCGT CTATGTTGTC TGTTACTATC	1320
5	TTGTTTTGT TTGTGGTTA CGGTTCTGT GTGTGTCTTG TGTGTCTCTT TGTGTTCAGA	1380
	CTTGGACTGA TGACTGACGA CTGTTTTAA GTTATGCCTT CTAAAATAAG CCTAAAAATC	1440
10	CTGTCAGATC CCTATGCTGA CCACCTCCTT TCAGATCAAC AGCTGCCCTT ACTCGAGCTC	1500
	AAGCTTCGAA TTCTGCAGTC GACGGTACCG CGGGCCCGGG ATCCACCGGT CGCCACCAG	1560
	GGTAAAGGAG AAGAACTTT CACTGGAGTT GTCCCAATT TTGTTGAATT AGATGGTGAT	1620
15	GTTAATGGGC ACAAAATTTTG TGTCAGTGGA GAGGGTGAAG GTGATGCAAC ATACGGAAAA	1680
	CTTACCCCTTA AATTTATTG CACTACTGGA AAAACTACCTG TTCCATGGCC AACACTTGTC	1740
20	ACTACTTTCT CTTATGGTGT TCAATGCTT TCAAGATAACC CAGATCATAT GAAACGGCAT	1800
	GACTTTTCA AGAGTGCAT GCCCGAAGGT TATGTACAGG AAAGAACTAT ATTTCCTCAA	1860
	GATGACGGGA ACTACAAGAC ACGTGCTGAA GTCAAGTTG AAGGTGATAC CCTTGTAA	1920
25	AGAACATCGAGT TAAAAGGTAT TGATTTAAA GAAGATGGAA ACATTCTTGG ACACAAATTG	1980
	GAATACAAC ATAACCTACA CAATGTATAC ATCATGGCAG ACAAAACAAAA GAATGGAAC	2040
30	AAAGTTAACT TCAAAATTAG ACACAACATT GAAGATGGAA GCCTTCAACT AGCAGACC	2100
	TATCAACAAA ATACTCCAAT TGGCGATGGC CCTGTCCTTT TACCAGACAA CCATTACCTG	2160
	TCCACACAAT CTGCCCTTTC GAAAGATCCC AACGAAAAGA GAGACCACAT GGTCTTCTT	2220
35	GAGTTGTAA CAGCTGCTGG GATTACACAT GGCAATGGATG AACTATACAA GTCCGGATCT	2280
	AGATAACTGT ATCGATGGAT CCGAAGGCGG GGACAGCAGT GCAGTGGTGG ACAGAAAGCA	2340
	AGTGATCTAG GCCAGCAGCC TCCCTAAAGG GACTTCAGCC CACAAAGCCA AACTTGTGGC	2400
40	TTTAATACAA GCTCTGTAAA TGGTAAAAAA AAAAAAGTCT ACACGGACAG CAGGTATGCT	2460
	CTTGCCACTG TACAGAGCAA TATACAGACA AAGAGAACTG TTGACATCTG CAGAGAAAGA	2520
45	CCTAAGATGC TGTGGCTAAA AGAAATCAGA TGGCAAATCT AACCGCCCGAG GCATCCTAAA	2580
	GAGCAATGAT CCTGACAGTC TGAAGACTAT CAAGTTATAG ACAAAATTAAG ACTGGTAAA	2640
	AAAACCTGT ATAAAATAGT AAAAAGTAA AAAAGAAAAC TAGTCCTCTC ATGAGAAGAC	2700
50	AGACCTGACA TCTACTGAAA AATAGACTTT ACTGGAAAAAA ATATGTGTAT GAATACCTTC	2760
	TAGTTTTGT GAACGTTCTC AAGATGGATA AAAGCTTTTC CTTGTAAAAC GAGACTGATC	2820
55	AGATAGTCAT CAAGAAGATT GTTAAAGAAA ATTTTCCAAG GTTCGGAGTG CCAAAAGCAA	2880
	TAGTGTCAAA TAATGGTCCT GCCTTGTG CCCAGGTAAG TCAGGGTGTG GCCAAGTATT	2940
	TAGAGGTCAA ATGAAAATTC CATTGTGTGT ACAGACCTCA GAGCTCAGGA AAGATAAAAA	3000
60	AGAATAAATA AAACTCTAAA CAGACCTTGA CAAAATTAAT CCTAGAGACT GGCACAGACT	3060
	TACTGGTAC TCCTTCCCCT TGCCCTATTT AGAACTGAGA ATACTCCCTC TTGATTGGT	3120
65	TTTACTCTTT TTAAGATCCT TTATGGGGCT CCTATGCCAT CACTGTCTTA AATGATGTGT	3180
	TTAAACCTAT GTTGTATAA TAATGATCTA TATGTAAAGT TAAAAGGCTT GCAGGTGGT	3240

	CAGAAAGAAG TCTGGTCACA ACTGGCTACA GTGAACAAAGC TGGGTACCCC AAGGACATCT	3300
5	TACCAGTTCC AGCCAGAGAT CTGATCTACG ATCCCCGGGT CGACCCGGGT CGACCCTGTG	3360
	GAATGTGTGT CAGTTAGGGT GTGGAAAGTC CCCAGGCTCC CCAGCAGGCA GAAGTATGCA	3420
	AAGCATGCAT CTCAATTAGT CAGCAACCAG GTGTGGAAAG TCCCCAGGCT CCCCCAGCAGG	3480
10	CAGAAGTATG CAAAGCATGC ATCTCAATT A GTCAGCAACC ATAGTCCCAGC CCCTAACTCC	3540
	GCCCCATCCCG CCCCTAACTC CGCCCAGTTC CGCCCATTCT CCGCCCCATG GCTGACTAAT	3600
15	TTTTTTTATT TATGCAGAGG CCGAGGCCGC CTCGGCCTCT GAGCTATTCC AGAAGTAGTG	3660
	AGGAGGCTTT TTTGGAGGCC TAGGCTTTG CAAAAAGCTT CACGCTGCCG CAAGCACTCA	3720
	GGGCGCAAGG GCTGCTAAAG GAAGCGGAAC ACGTAGAAAG CCAGTCCGCA GAAACGGTGC	3780
20	TGACCCCCGA TGAATGTCAG CTACTGGCT ATCTGGACAA GGGAAAACGC AAGCGCAAAG	3840
	AGAAAAGCAGG TAGCTTGCAG TGGGCTTACA TGGCGATAGC TAGACTGGC GGTTTTATGG	3900
25	ACAGCAAGCG AACCGGAATT GCCAGCTGGG GCGCCCTCTG GTAAGGTTGG GAAGCCCTGC	3960
	AAAGTAAACT GGATGGCTTT CTTGCCGCCA AGGATCTGAT GGCGCAGGGG ATCAAGATCT	4020
	GATCAAGAGA CAGGATGAGG ATCGTTCGC ATGATTGAAC AAGATGGATT GCACGCAGGT	4080
30	TCTCCGGCCG CTTGGGTGGA GAGGCTATTG GGCTATGACT GGGCACAAACA GACAATCGGC	4140
	TGCTCTGATG CCGCCGTGTT CCGGCTGTCA GCGCAGGGC GCCCCGTTCT TTTGTCAAG	4200
35	ACCGACCTGT CCGGTGCCCT GAATGAACTG CAGGACGAGG CAGCGCGGCT ATCGTGGCTG	4260
	GCCACGACGG GCGTTCTTG CGCAGCTGTG CTCGACGTG TCACTGAAGC GGGAAAGGGAC	4320
	TGGCTGCTAT TGGGCGAAGT GCGGGGGCAG GATCTCCTGT CATCTCACCT TGCTCCTGCC	4380
	GAGAAAGTAT CCATCATGGC TGATGCAATG CGGGGGCTGC ATACGTTGA TCCGGCTACC	4440
40	TGCCCATTG ACCACCAAGC GAAACATCGC ATCGAGCGAG CACGTACTCG GATGGAAGCC	4500
	GGTCTTGTG ATCAGGATGA TCTGGACGAA GAGCATCAGG GGCTCGCGCC AGCCGAACGT	4560
45	TTCGCCAGGC TCAAGGCGCG CATGCCGAC GGCAGGATC TCGTCGTGAC CCATGGCGAT	4620
	GCCTGCTTGC CGAATATCAT GGTGGAAAAT GGCCGCTTT CTGGATTGAT CGACTGTGGC	4680
	CGGCTGGGTG TGGCGGACCG CTATCAGGAC ATAGCGTTGG CTACCCGTGA TATTGCTGAA	4740
50	GAGCTTGGCG GCGAATGGGC TGACCGCTTC CTCGTGCTT ACGGTATCGC CGCTCCCGAT	4800
	TCGCAGCGCA TCGCCTTCTA TCGCCTTCTT GACGAGTTCT TCTGAGCGGG ACTCTGGGT	4860
55	TCGAAATGAC CGACCAAGCG ACGCCCAACC TGCCATCACG AGATTCGAT TCCACCGCCG	4920
	CCTTCTATGA AAGGTTGGGC TTCGGAATCG TTTCCGGGA CGGAATTGCGT AATCTGCTGC	4980
	TTGCAAACAA AAAAACCAACC GCTACCAGCG GTGGTTGTT TGCCGGATCA AGAGCTACCA	5040
60	ACTCTTTTC CGAAGGTAAC TGGCTTCAGC AGAGCGCAGA TACCAAATAC TGTCCCTCTA	5100
	GTGTAGCCGT AGTTAGGCCA CCACCTCAAG AACTCTGTAG CACCGCCTAC ATACCTCGCT	5160
65	CTGCTAATCC TGTTACCAGT GGCTGCTGCC AGTGGCGATA AGTCGTGTCT TACCGGGTTG	5220

	GACTCAAGAC GATAAGTTACC GGATAAGGCG CAGCGGTGG GCTGAACGGG GGGTTCGTGC	5280
	ACACAGCCCA GCTTGGAGCG AACGACCTAC ACCGAACCTGA GATACCTACA GCGTGAGCAT	5340
5	TGAGAAAGCG CCACGCTTCC CGAAGGGAGA AAGGCGGACA GGTATCCGGT AAGCggcagg	5400
	GTCGGAACAG GAGAGCGCAC GAGGGAGCTT CCAGGGGGAA ACGCCTGGTA TCTTTATAGT	5460
10	CCTGTCGGGT TTGCCCCACCT CTGACTTGAG CGTCGATTTC TGTGATGCTC GTCAGGGGG	5520
	CGGAGCCTAT GGAAAAACGC CAGCAACGCC GAGATGCCGC GCCTCGAGTA CACCTGCGTC	5580
	ATGCTGAGAC CCTCAAGCCT CACTAAAAGG GTCCCTGCCT AGTTCTGTT ACTAATCTGC	5640
15	CTTATTCTGT TTTTGTTCCT ATGTTAAAGA TAGAGTAAAT GCAGTATTCT CCACATAGAG	5700
	ATATAGACTT CTGAAATTCT AAGATTAGAA TTATTTACAA GAAGAAGTGG GGAA	5754

## (2) INFORMATION FOR SEQ ID NO:20:

- 20 (i) SEQUENCE CHARACTERISTICS:  
 (A) LENGTH: 4958 base pairs  
 (B) TYPE: nucleic acid  
 (C) STRANDEDNESS: single  
 (D) TOPOLOGY: linear
- 25 (ii) MOLECULE TYPE: DNA (genomic)

## 30 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:20:

	AGGCGGGGAC AGCAGTGCAG TGGTGGACAG AAAGCAAGTG ATCTAGGCCA GCAGCCTCCC	60
35	TAAAGGGACT TCAGCCCACA AAGCCAAACT TGTGGCTTTA ATACAAGCTC TGTAAATGGT	120
	AAAAAAAAAA AAGTCTACAC GGACAGCAGG TATGCTCTTG CCACGTACA GAGCAATATA	180
40	CAGACAAAGA GAACTGTTGA CATCTGCAGA GAAAGACCTA AGATGCTGTG GCTAAAAGAA	240
	ATCAGATGGC AAATCTAACCC GCCCAGGCAT CCTAAAGAGC AATGATCCTG ACAGTCTGAA	300
	GACTATCAAG TTATAGACAA ATTAAGACTG GTAAAAAAAAA CCCTGTATAA AATAGTAAAA	360
45	ACTGAAAAAA GAAAAGTAGT CCTCTCATGA GAAGACAGAC CTGACATCTA CTGAAAAATA	420
	GACTTTACTG GAAAAAATAT GTGTATGAAT ACCTTCTAGT TTTTGTGAAC GTTCTCAAGA	480
50	TGGATAAAAG CTTTCCTTG TAAAACGAGA CTGATCAGAT AGTCATCAAG AAGATTGTTA	540
	AAGAAAATT TCCAAGGTTC GGAGTGCCAA AAGCAATAGT GTCAGATAAT GGTCCTGCCT	600
55	TTGTTGCCCA GGTAAGTCAG GGTGTGGCCA AGTATTTAGA GGTCAAATGA AAATTCCATT	660
	GTGTGTACAG ACCTCAGAGC TCAGGAAAGA TAAAAAAGAA TAAATAAAAC TCTAAACAGA	720
	CCTTGACAAA ATTAATCCTA GAGACTGGCA CAGACTTAAT TGGTACTCCT TCCCCTGCC	780
60	CTATTTAGAA CTGAGAATAC TCCCTCTTGA TTCGGTTTA CTCTTTAA GATCCTTTAT	840
	GGGGCTCCTA TGCCATCACT GTCTTAAATG ATGTGTTAA ACCTATGTTG TTATAATAAT	900
	GATCTATATG TTAAGTTAAA AGGCTTGCAG GTGGTGCAGA AAGAAGTCTG GTCACAAC	960
65	GCTACAGTGA ACAAGCTGGG TACCCCAAGG ACATCTTACC AGTCCAGCC AGAGATCTGA	1020

	TCTACGATCC CCGGGTCGAC CCGGGTGAC CCTGTGGAAT GTGTGTCAGT TAGGGTGTGG	1080
5	AAAGTCCCCA GGCTCCCCAG CAGGCAGAAG TATGCAAAGC ATGCATCTCA ATTAGTCAGC	1140
	AACCAGGTGT GGAAAGTCCC CAGGCTCCCC AGCAGGCAGA AGTATGCAAA GCATGCATCT	1200
	CAATTAGTCA GCAACCATAG TCCC GCCCCT AACTCCGCC ATCCGCCCT TAACTCCGCC	1260
10	CAGTTCCGCC CATTCTCCGC CCCATGGCTG ACTAATTTT TTTATTTATG CAGAGGCCGA	1320
	GGCCGCCCTCG GCCTCTGAGC TATTCCAGAA GTAGTGAGGA GGCTTTTTG GAGGCCTAGG	1380
15	CTTTGCAAA AAGCTTCACG CTGCCCAAG CACTCAGGGC GCAAGGGCTG CTAAAGGAAG	1440
	CGGAACACGT AGAAAGCCAG TCCGCAGAAA CGGTGCTGAC CCCGGATGAA TGTCAGCTAC	1500
	TGGGCTATCT GGACAAGGGAA AAACGCAAGC GCAAAGAGAA AGCAGGTAGC TTGCAGTGGG	1560
20	CTTACATGGC GATAGCTAGA CTGGCGGTT TTATGGACAG CAAGCGAAC GGAATTGCCA	1620
	GCTGGGGCGC CCTCTGGTAA GGTTGGGAAG CCCTGCAAAG TAAACTGGAT GGCTTTCTTG	1680
25	CCGCCAAGGA TCTGATGGCG CAGGGGATCA AGATCTGATC AAGAGACAGG ATGAGGATCG	1740
	TTTCGCATGA TTGAACAAGA TGGATTGCAC GCAGGTTCTC CGGCCGCTTG GGTGGAGAGG	1800
	CTATTGGCT ATGACTGGGC ACAACAGACA ATCGGCTGCT CTGATGCCGC CGTGTCCGG	1860
30	CTGTCAGCGC AGGGGCGCCC GGTTCTTTT GTCAAGACCG ACCTGTCCGG TGCCCTGAAT	1920
	GAACCTGCAGG ACGAGGCAGC GCGGCTATCG TGGCTGGCCA CGACGGCGT TCCTTGCAGCA	1980
35	GCTGTGCTCG ACGTTGTCAC TGAAGCGGGA AGGGACTGGC TGCTATTGGG CGAAGTGCAG	2040
	GGGCAGGATC TCCTGTCATC TCACCTTGCT CCTGCCGAGA AAGTATCCAT CATGGCTGAT	2100
	GCAATGCGGC GGCTGCATAC GCTTGATCCG GCTACCTGCC CATTGACCA CCAAGCGAAA	2160
40	CATCGCATCG AGCGAGCAGC TACTCGGATG GAAGCCGGTC TTGTCGATCA GGATGATCTG	2220
	GACGAAGAGC ATCAGGGGCT CGCGCCAGCC GAACTGTTCG CCAGGCTCAA GGCGCGCATG	2280
	CCCGACGGCG AGGATCTCGT CGTGACCCAT GGCGATGCCT GCTTGCCGAA TATCATGGTG	2340
45	GAAAATGGCC GCTTTCTGG ATTGATCGAC TGTGGCCGGC TGGGTGTGGC GGACCGCTAT	2400
	CAGGACATAG CGTTGGCTAC CCGTGATATT GCTGAAGAGC TTGGCGGCGA ATGGGCTGAC	2460
	CGCTTCTCG TGCTTTACGG TATGCCGCT CCCGATTCCG AGCCGATCGC CTTCTATCGC	2520
50	CTTCTTGACG AGTTCTCTG AGCGGGACTC TGGGGTTCGA AATGACCGAC CAAGCGACGC	2580
	CCAACCTGCC ATCACGAGAT TTGATTCGA CCGCCGCCTT CTATGAAAGG TTGGGCTTCG	2640
55	GAATCGTTT CCGGGACGGA ATTGTAATC TGCTGCTTCG AAACAAAAAA ACCACCGCTA	2700
	CCAGCGGTGG TTTGTTGCC GGATCAAGAG CTACCAACTC TTTTCCGAA GGTAACTGGC	2760
60	TTCAGCAGAG CGCAGATAACC AAATACTGTC CTTCTAGTGT AGCCGTAGTT AGGCCACCAC	2820
	TTCAAGAACT CTGTAGCACC GCCTACATAC CTCGCTCTGC TAATCCTGTT ACCAGTGGCT	2880
	GCTGCCAGTG GCGATAAGTC GTGTCTTACG GGGTTGGACT CAAGACGATA GTTACCGGAT	2940
65	AAGGCGCAGC GGTCGGGCTG AACGGGGGGT TCGTGCACAC AGCCCAGCTT GGAGCGAACG	3000

	ACCTACACCG AACTGAGATA CCTACAGCGT GAGCATTGAG AAAGGCCAC GCTTCCGAA	3060
	GGGAGAAAGG CGGACAGGTA TCCGGAAGC GGCAGGGTCG GAACAGGAGA GCGCACGAGG	3120
5	GAGCTTCCAG GGGGAAACGC CTGGTATCTT TATAGTCCTG TCGGGTTCG CCACCTCTGA	3180
	CTTGAGCGTC GATTTTGAG ATGCTCGTCA GGGGGCGGA GCCTATGGAA AACGCCAGC	3240
10	AACGCCAGA TGCGCCGCCT CGAGTACACC TGCGTCATGC TGAGACCCTC AAGCCTCACT	3300
	AAAAGGGTCC CTGCCTAGTT CTGTTACTA ATCTGCCTTA TTCTGTTTT GTTCCCATGT	3360
	TAAAGATAGA GTAAATGCAG TATTCTCCAC ATAGAGATAT AGACTTCTGA AATTCTAAGA	3420
15	TTAGAATTAT TTACAAGAAC AAGTGGGAA TGAAGAATAA AAAATTACTG GCCTCTTGTG	3480
	AGAACATGAA CTTTCACCTC GGAGCCCACC CCCTCCCATC TGGAAAACAT ACTTGAGAAA	3540
20	AACATTTCT GGAACAAACCA CAGAATGTTT CAACAGGCCA GATGTATTGC CAAACACAGG	3600
	ATATGACTCT TTGGTTGAGT AAATTGTGG TTGTTAAACT TCCCCTATT CCTCCCCATT	3660
	CCCCCTCCCA GTTTGTGGTT TTTCTTTA AAAGCTTGTG AAAAATTGA GTCGTCGTG	3720
25	AGACTCCTCT ACCCTGTGCA AAGGTGTATG AGTTTCGACC CCAGAGCTCT GTGTGCTTTC	3780
	TGTTGCTGCT TTATTCGAC CCCAGAGCTC TGGTCTGTGT GCTTCATGT CGCTGCTTTA	3840
30	TTAAATCTTA CCTTCTACAT TTTATGTATG GTCTCAGTGT CTTCTGGGT ACGCGGCTGT	3900
	CCCGGGACTT GAGTGTCTGA GTGAGGGTCT TCCCTCGAGG GTCTTCATT TGGTACATGG	3960
	GCCGGGAATT CGAGAATCTT TCATTGGTG CATTGGCCGG GAATTGAAA ATCTTCATT	4020
	TGGTGCATTG GCCGGGAAAC AGCGCGACCA CCCAGAGGTC CTAGACCCAC TTAGAGGTAA	4080
35	GATTCTTGT TCTGTTTGG TCTGATGTCT GTGTTCTGAT GTCTGTGTTG TGTTCTAAG	4140
	TCTGGTGCAG TCGCAGTTTC AGTTTGCAGG ACGCTCAGTG AGACCGCGCT CCGAGAGGGA	4200
40	GTGCGGGGTG GATAAGGATA GACGTGTCCA GGTGTCCACC GTCCGTCGCG CCTGGGAGAC	4260
	GTCCCAGGAG GAACAGGGGA GGATCAGGGA CGCCTGGTGG ACCCCTTGA AGGCCAAGAG	4320
45	ACCATTGGG GTTGCAGAGAT CGTGGGTTCG AGTCCCACCT CGTCCCAGT TGCGAGATCG	4380
	TGGGTTCGAG TCCCACCTCG TGTTTGTG CGAGATCGT GGTCGAGTC CCACCTCGCG	4440
	TCTGGTCACG GGATCGTGGG TTCGAGTCCC ACCTCGTGTG TTGTTGCGAG ATCGTGGGTT	4500
50	CGAGTCCCAC CTCGCGTCTG GTCACGGGAT CGTGGGTTCG AGTCCCACCT CGTGCAGAGG	4560
	GTCTCAATTG GCCGGCCTTA GAGAGGCCAT CTGATTCTTC TGGTTCTCT TTTGTCTTA	4620
55	GTCTCGTGTGTC CGCTCTTGTG GTGACTACTG TTTTCTAAA AATGGGACAA TCTGTGTCCA	4680
	CTCCCCTTTC TCTGACTCTG GTTCTGTCGC TTGGTAATT TGTTGTTA CGTTTGTGTT	4740
	TGTGAGTCGT CTATGTTGTC TGTTACTATC TTGTTTTGT TTGTTGTTA CGGTTCTGT	4800
60	GTGTGTCTTG TGTGTCTCTT TGTGTTCAGA CTTGGACTGA TGACTGACGA CTGTTTTAA	4860
	GTTATGCCTT CTAAAATAAG CCTAAAAATC CTGTCAGATC CCTATGCTGA CCACCTCCTT	4920
65	TCAGATCAAC AGCTGCCCTT ACGTATCGAT GGATCCGA	4958

(2) INFORMATION FOR SEQ ID NO:21:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 7080 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

10

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:21:

15	GAATACAAGC TTGCATGCCT GCAGGTCGAC TCTAGAGGAT CTTGAAGAAT AAAAAATTAC TGGCCTCTTG TGAGAACATG AACTTCAC CCGGAGCCCA CCCCTCCC TCTGGAAAAC	60 120
20	ATACTTGAGA AAAACATTCT CTGGAACAAAC CACAGAATGT TTCAACAGGC CAGATGTATT GCCAAACACA GGATATGACT CTTTGGTTGA GTAAATTGT GGTTGTTAAA CTTCCCTAT	180 240
25	TCCCTCCCCA TTCCCCCTCC CAGTTGTGG TTTTTCCCTT TAAAAGCTTG TGAAAAATT GAGTCGTCGT CGAGACTCCT CTACCCGTG CAAAGGTGTA TGAGTTTCGA CCCCAGAGCT	300 360
30	CTGTGTGCTT TCTGTTGCTG CTTTATTTCG ACCCCAGAGC TCTGGTCTGT GTGCTTCAT GTCGCTGCTT TATTAATCT TACCTTCTAC ATTTTATGTA TGGTCTCAGT GTCTTCTTGG	420 480
35	GTACGCGGCT GTCCCGGGAC TTGAGTGTCT GAGTGAGGGT CTTCCCTCGA GGGTCTTC TTTGGTACAT GGGCCGGGAA TTCGAGAACATC TTTCATTTGG TGCATTGGCC GGGATTGGA	540 600
40	AAATCTTCAT TTTGGTGCAT TGGCCGGGAA ACAGCGCGAC CACCCAGAGG TCCTAGACCC ACTTAGAGGT AAGATTCTTT GTTCTGTTTT GGTCTGATGT CTGTGTTCTG ATGTCTGTGT	660 720
45	TCTGTTCTA AGTCTGGTGC GATCGCAGTT TCAGTTTGC GGACGCTCAG TGAGACC CTCCGAGAGG GAGTGCAGGG TGGATAAGGA TAGACGTGTC CAGGTGTCCA CCCTCCGTT GCCCTGGGAG ACGTCCCAGG AGGAACAGGG GAGGATCAGG GACGCCTGGT GGACCC 900	780 840
50	960 GAAGGCCAAG AGACCATTG GGGTTGCGAG ATCGTGGGTT CGAGTCCCAC CATCGATGGT GCAGAGGGTC TCAATTGGCC GGCCTTAGAA TTACGGATCT AGCATGATTG AACAAAGATGG ATTGCACGCA GGTCTCCGG CCGCTGGGT GGAGAGGCTA TTGGCTATG ACTGGGCACA ACAGACAATC GGCTGCTCTG ATGCCGCCGT GTTCCGGCTG TCAGCGCAGG GGCGCCCGGT TCTTTTGTC AAGACCGACC TGTCCGGTGC CCTGAATGAA CTGCAGGACG AGGCAGCGCG	1020 1080 1140 1200
55	GCTATCGTGG CTGGCCACGA CGGGCGTTCC TTGCGCAGCT GTGCTCGACG TTGTC AGCGGGAAAG GACTGGCTGC TATTGGCGA AGTGCCGGGG CAGGATCTCC TGTCATCTCA CCTTGCTCCT GCCGAGAAAG TATCCATCAT GGCTGATGCA ATGCGCGGC TGCATACGCT	1260 1320 1380
60	TGATCCGGCT ACCTGCCAT TCGACCACCA AGCGAAACAT CGCATCGAGC GAGCAC TCGGATGGAA GCCGGTCTTG TCGATCAGGA TGATCTGGAC GAAGAGCATC AGGGGCTCGC	1440 1500
65	GCCAGCGCAA CTGTTGCCA GGCTCAAGGC GCGCATGCC GACGGCGAGG ATCTCGTCGT	1560

	GACCCATGGC GATGCCTGCT TGCCGAATAT CATGGTGGAA AATGCCGCT TTTCTGGATT	1620
	CATCGACTGT GGCGGGCTGG GTGTGGCGGA CCGCTATCAG GACATAGCGT TGGCTACCG	1680
5	TGATATTGCT GAAGAGCTTG GCGGCGAATG GGCTGACCAG TTCCCTCGTGC TTTACGGTAT	1740
	CGCCGCTCCC GATT CGCAGC GCAT CGCCTT CTAT CGCCTT CTTGACGAGT TCTTCTGAGC	1800
10	GGGACTCTGG GTT CGTAAT GACCGACCAA GCGACGCCA ACCTGCCATC ACGAGATTTC	1860
	GATTCCACCG CCGCCTCTA TGAAAGGTTG GGCTTCGGAG TTAGCTTGTT TCTTTACTGT	1920
	TTGTCAATTG TATTATTTCA ATACAGAACAA ATAGCTTCTA TAACTGAAAT ATATTTGCTA	1980
15	TTGTATATTA TGATTGTCCC TCGAACCATG AACACTCCTC CAGCTGAATT TCACAATTCC	2040
	TCTGTCATCT GCCAGGCCAT TAAGTTATTC ATGGAAGATC TTTGAGGAAC ACTGCAAGTT	2100
20	CATATCATAA ACACATTTGA AATTGAGTAT TGTTTGACAT TGTATGGAGC TATGTTTG	2160
	TGTATCCTCA GAAAAAAAAGT TTGTTATAAA GCATTCACAC CCATAAAAAG ATAGATTAA	2220
	ATATTCCAGC TATAGGAAAG AAAGTGCCTC TGCTCTTCAC TCTAGTCTCA GTTGGCTCCT	2280
25	TCACATGCAT GCTTCTTTAT TTCTCCTATT TTGTCAAGAA AATAATAGGT CACGTCTTGT	2340
	TCTCACTTAT GTCCTGCCTA GCATGGCTCA GATGCACGTT GTAGATACAA GAAGGATCAA	2400
30	ATGAAACAGA CTTCTGGTCT GTTACTACAA CCATAGTAAT AAGCACACTA ACTAATAATT	2460
	GCTAATTATG TTTTCCATCT CTAAGGTTCC CACATTTTC TGTTTCTTA AAGATCCCAT	2520
	TATCTGGTTG TAACTGAAGC TCAATGGAAC ATGAGCAATA TTTCCCAGTC TTCTCTCCC	2580
35	TCCAACAGTC CTGATGGATT AGCAGAACAG GCAGAAAACA CATTGTTACC CAGAATTAAA	2640
	AACTAATATT TGCTCTCCAT TCAATCCAAA ATGGACCTAT TGAAACTAAA ATCTAACCCA	2700
	ATCCCATTAA ATGATTCTA TGGCGTCAAA GGTCAAACCTT CTGAAGGGAA CCTGTGGGTG	2760
40	GGTCACAATT CAGGCTATAT ATTCCCCAGG GCTCAGCCAG TGTCTGTACA TACACAACGG	2820
	ATCCTGTGGA CAGCTCACCT AGCTGCAATG GCTACAGGCT CCCGGACGTC CCTGCTCCTG	2880
	GCTTTGGCC TGCTCTGCCT GCCCTGGCTT CAAGAGGGCA GTGCCCTCCC AACCATTC	2940
45	TTATCCAGGC TTTTGACAA CGCTATGCTC CGCGCCCATC GTCTGCACCA GCTGGCCTTT	3000
	GACACCTACC AGGAGTTGA AGAAGCCTAT ATCCCAAAGG AACAGAAGTA TTCATTCTG	3060
50	CAGAACCCCC AGACCTCCCT CTGTTCTCA GAGTCTATTG CGACACCCTC CAACAGGGAG	3120
	GAAACACAAC AGAAATCCAA CCTAGAGCTG CTCCGCATCT CCCTGCTGCT CATCCAGTCG	3180
55	TGGCTGGAGC CCGTGCAGTT CCTCAGGAGT GTCTCGCCA ACAGCCTGGT GTACGGCGCC	3240
	TCTGACAGCA ACGTCTATGA CCTCCTAAAG GACCTAGAGG AAGGCATCCA AACGCTGATG	3300
	GGGAGGCTGG AAGATGGCAG CCCCCGGACT GGGCAGATCT TCAAGCAGAC CTACAGCAAG	3360
60	TTCGACACAA ACTCACACAA CGATGACGCA CTACTCAAGA ACTACGGCT GCTCTACTGC	3420
	TTCAGGAAGG ACATGGACAA GGTGAGACA TTCCCTGGCA TCGTGCAGTG CCGCTCTGTG	3480
65	GAGGGCAGCT GTGGCTTCTA GCTGCCCGGG TGGCATCCTG TGACCCCTCC CCAGTGCCTC	3540
	TCCTGCCCT GGAAGTTGCC ACTCCAGTGC CCACCAGCCT TGTCTTAATA AAATTAAGTT	3600

	GCATCAAAAA AAAAAAAAAG CTAGCGGCCG CTAGACTTCT GAAATTCTAA GATTAGAATT	3660
5	ATTTACAAGA AGAAGTGGGG AATGAAGAAT AAAAAATTAC TGGCCTCTTG TGAGAACATG	3720
	AACTTTCACC TCGGAGCCA CCCCCTCCCA TCTGGAAAAC ATACTTGAGA AAAACATTT	3780
	CTGGAACAAC CACAGAACATGT TTCAACAGGC CAGATGTATT GCCAAACACA GGATATGACT	3840
10	CTTGGTGA GTAAATTGT GGTTGTTAAA CTTCCCTAT TCCCTCCCCA TTCCCCCTCC	3900
	CAGTTTGTGG TTTTTTCCTT TAAAAGCTTG TGAAAAATTG GAGTCGTCGT CGAGACTCCT	3960
15	CTACCCCTGTG CAAAGGTGTA TGAGTTCGA CCCCAGAGCT CTGTGTGCTT TCTGTTGCTG	4020
	CTTTATTTCG ACCCCAGAGC TCTGGTCTGT GTGCTTCAT GTCGCTGCTT TATTAAATCT	4080
	TACCTCTAC ATTTATGTA TGGTCTCAGT GTCTTCTTGG GTACCGGGCT GTCCCGGGAC	4140
20	TTGAGTGTCT GAGTGAGGGT CTTCCCTCGA GGGTCTTCA TTTGGTACAT GGGCCGGGAA	4200
	TTCGAGAACATC TTTCATTGG TGCATTGCC GGGAAATTCGA AAATCTTCA GATCCCCGGG	4260
25	TACCGAGCTC GAATTCCGGT CTCCCTATAG TGAGTCGTAT TAATTCGAT AAGCCAGCTG	4320
	CATTAATGAA TCGGCCAACG CGCGGGGAGA GGCGGTTTGC GTATTGGCG CTCTCCGCT	4380
	TCCTCGCTCA CTGACTCGCT GCGCTCGGTC GTTCGGCTGC GGCGAGCGGT ATCAGCTCAC	4440
30	TCAAAGGCAGG TAATACGGTT ATCCACAGAA TCAGGGGATA ACGCAGGAAA GAACATGTGA	4500
	GCAAAAGGCC AGCAAAAGGC CAGGAACCGT AAAAAGGCCG CGTTGCTGGC GTTTTCCAT	4560
	AGGCTCCGCC CCCCTGACGA GCATCACAAA AATCGACGCT CAAGTCAGAG GTGGCGAAC	4620
35	CCGACAGGAC TATAAAGATA CCAGGCCTTT CCCCCTGGAA GCTCCCTCGT GCGCTCTCCT	4680
	GTTCCGACCC TGCCGCTTAC CGGATACCTG TCCGCCTTTC TCCCTTCGGG AAGCGTGGCG	4740
40	CTTTCTCATA GCTCACGCTG TAGGTATCTC AGTCGGTGT AGGTGTTCG CTCCAAGCTG	4800
	GGCTGTGTGC ACGAACCCCC CGTTCAGCCC GACCGCTGCG CCTTATCCGG TAACTATCGT	4860
	CTTGAGTCCA ACCCGGTAAG ACACGACTTA TCGCCACTGG CAGCAGCCAC TGGTAACAGG	4920
45	ATTAGCAGAG CGAGGTATGT AGGCGGTGCT ACAGAGTTCT TGAAGTGGTG GCCTAACTAC	4980
	GGCTACACTA GAAGGACAGT ATTTGGTATC TGCGCTCTGC TGAAGCCAGT TACCTTCGGA	5040
	AAAAGAGTTG GTAGCTCTTG ATCCGGCAAA CAAACCACCG CTGGTAGCGG TGGTTTTTT	5100
50	GTTTGCAAGC AGCAGATTAC GCGCAGAAAA AAAGGATCTC AAGAAGATCC TTTGATCTTT	5160
	TCTACGGGGT CTGACGCTCA GTGGAACGAA AACTCACGTT AAGGGATTTT GGTCATGAGA	5220
55	TTATCAAAAA GGATCTTCAC CTAGATCCTT TTAAATTAAA AATGAAGTTT TAAATCAATC	5280
	TAAAGTATAT ATGAGTAAAC TTGGTCTGAC AGTTACCAAT GCTTAATCAG TGAGGCACCT	5340
60	ATCTCAGCGA TCTGTCTATT TCGTTCATCC ATAGTTGCCT GACTCCCCGT CGTGTAGATA	5400
	ACTACGATAAC GGGAGGGCTT ACCATCTGGC CCCAGTGCTG CAATGATACC GCGAGACCCA	5460
	CGCTCACCGG CTCCAGATTG ATCAGCAATA AACCAAGCCAG CCGGAAGGGC CGAGCGCAGA	5520
65	AGTGGTCTTG CAACTTTATC CGCCTCCATC CAGTCTATTA ATTGTTGCCG GGAAGCTAGA	5580

	GTAAGTAGTT CGCCAGTTAA TAGTTGCGC AACGTTGTTG CCATTGCTAC AGGCATCGTG	5640
	GTGTCACGCT CGTCGTTGG TATGGCTTCA TTCAGCTCCG GTTCCCAACG ATCAAGGCGA	5700
5	GTTACATGAT CCCCCATGTT GTGAAAAAAA GCGGTTAGCT CCTTCGGTCC TCCGATCGTT	5760
	GTCAGAAGTA AGTTGGCCGC AGTGTATCA CTCATGGTTA TGGCAGCACT GCATAATTCT	5820
10	CTTACTGTCA TGCCATCCGT AAGATGCTTT TCTGTGACTG GTGAGTACTC AACCAAAGTCA	5880
	TTCTGAGAAT AGTGTATGCG GCGACCGAGT TGCTCTTGCC CGGCGTCAAT ACGGGATAAT	5940
	ACCGCGCCAC ATAGCAGAAC TTTAAAAGTG CTCATCATTG GAAAACGTTT TTGGGGCGA	6000
15	AAACTCTCAA GGATCTTACC GCTGTTGAGA TCCAGTTCGA TGTAACCCAC TCGTGCACCC	6060
	AACTGATCTT CAGCATCTTT TACTTTCACCC AGCGTTCTG GGTGAGCAAA AACAGGAAGG	6120
20	CAAAATGCCG CAAAAAAGGG AATAAGGGCG ACACGGAAAT GTTGAATACT CATACTCTC	6180
	CTTTTTCAAT ATTATTGAAG CATTTATCAG GGTTATTGTC TCATGAGCGG ATACATATTT	6240
	GAATGTATTT AGAAAAATAA ACAAAATAGGG GTTCCCGCGA CATTCCCCG AAAAGTGCCA	6300
25	CCTGACGTCT AAGAAACCAT TATTATCATG ACATTAACCT ATAAAAAATAG GCGTATCACG	6360
	AGGCCCTTTC GTCTCGCGCG TTTCGGTGAT GACGGTGAAA ACCTCTGACA CATGCAGCTC	6420
30	CCGGAGACGG TCACAGCTTG TCTGTAAGCG GATGCCGGGA GCAGACAAGC CCGTCAGGGC	6480
	GCGTCAGCGG GTGTTGGCGG GTGTCGGGGC TGGCTTAACT ATGCGGCATC AGAGCAGATT	6540
	GTACTGAGAG TGCACCATAT CGACGCTCTC CCTTATGCGA CTCCTGCATT AGGAAGCAGC	6600
35	CCAGTAGTAG GTTGAGGCCG TTGAGCACCG CCGCCGCAAG GAATGGTGCA AGGAGATGGC	6660
	GCCCCAACAGT CCCCCGGCCA CGGGGCCTGC CACCATACCC ACGCCGAAAC AAGCGCTCAT	6720
	GAGCCCCGAAG TGGCGAGCCC GATCTTCCCC ATCGGTGATG TCGGGCATAT AGGCGCCAGC	6780
40	AACCGCACCT GTGGCGCCGG TGATGCCGGC CACGATGCGT CCGGCGTAGA GGATCTGGCT	6840
	AGCGATGACC CTGCTGATTG GTTCGCTGAC CATTCCGGG GTGCGGAACG GCGTTACCAAG	6900
	AAACTCAGAA GGTCGTCCA ACCAAACCGA CTCTGACGGC AGTTTACGAG AGAGATGATA	6960
45	GGGTCTGCTT CAGTAAGCCA GATGCTACAC AATTAGGCTT GTACATATTG TCGTTAGAAC	7020
	GGGGCTACAA TTAATACATA ACCTTATGTA TCATACACAT ACGATTAGG TGACACTATA	7080

(2) INFORMATION FOR SEQ ID NO:22:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 6795 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:22:

AATGAAAAGAC CCCACCTGTA GGTGTTGGCAA GCTAGCTTAA GTAACGCCAT TTTGCAAGGC

	ATGGAAAAAT ACATAACTGA GAATAGAGAA GTTCAGATCA AGGTCAAGGAA CAGATGGAAC	120
5	AGCTGAATAT GGGCCAAACA GGATATCTGT GGTAAGCAGT TCCTGCCCG GCTCAGGGCC	180
	AAGAACAGAT GGAACAGCTG AATATGGGCC AAACAGGATA TCTGTGGTAA GCAGTTCTG	240
	CCCCGGCTCA GGGCCAAGAA CAGATGGTCC CCAGATGCCGG TCCAGCCCTC AGCAGTTCT	300
10	AGAGAACCAT CAGATGTTTC CAGGGTGCCC CAAGGACCTG AAATGACCCT GTGCCTTATT	360
	TGAACTAACCC AATCAGTTCG CTTCTCGCTT CTGTTCGCGC GCTTCTGCTC CCCGAGCTCA	420
15	ATAAAAGAGC CCACAAACCCC TCACTCGGGG CGCCAGTCCT CCGATTGACT GAGTCGCCCG	480
	GGTACCCGTG TATCCAATAA ACCCTCTTGC AGTTGCATCC GACTTGTGGT CTCGCTGTT	540
	CTTGGGAGGG TCTCCTCTGA GTGATTGACT ACCCGTCAGC GGGGGTCTTT CATTGGGGG	600
20	CTCGTCCGGG ATCGGGAGAC CCCTGCCAG GGACCACCGA CCCACCACCG GGAGGTAAGC	660
	TGGCCAGCAA CTTATCTGTG TCTGTCCGAT TGTCTAGTGT CTATGACTGA TTTTATGCGC	720
25	CTGCGTCGGT ACTAGTTAGC TAACTAGCTC TGTATCTGGC GGACCCGTGG TGGAACGTGAC	780
	GAGTTCGGAA CACCCGGCCG CAACCTGGG AGACGTCCCA GGAGGAACAG GGGAGGATCA	840
	GGGACGCCTG GTGGACCCCT TTGAAGGCCA AGAGACCATT TGGGGTTGCG AGATCGTGGG	900
30	TTCGAGTCCC ACCTCGTGCC CAGTTGCGAG ATCGTGGGTT CGAGTCCCAC CTCGTGTTT	960
	GTTGCGAGAT CGTGGGTTCG AGTCCCACCT CGCGTCTGGT CACGGGATCG TGGGTTCGAG	1020
	TCCCCACCTCG TGTTTGTGTT CGAGATCGTG GGTCGAGTC CCACCTCGCG TCTGGTCACG	1080
35	GGATCGTGGG TTCGAGTCCC ACCTCGTGCA GAGGGTCTCA ATTGGCCGGC CTTAGAGAGG	1140
	CCATCTGATT CTTCTGGTTT CTCTTTGTGTT CTTAGTCTCG TGTCCGCTCT TGTTGTGACT	1200
40	ACTGTTTTTC TAAAATGGG ACAATCTGTG TCCACTCCCC TTTCTCTGAC TCTGGTTCTG	1260
	TCGCTTGGTA ATTTGTTTG TTTACGTTTG TTTTGTGAG TCGTCTATGT TGTCTGTTAC	1320
	TATCTTGTGTT TTGTTTGTGG TTTACGGTTT CTGTGTGTGT CTTGTGTGTC TCTTTGTGTT	1380
45	CAGACTTGGG CTGATGACTG ACGACTGTTT TTAAGTTATG CCTTCTAAAA TAAGCCTAAA	1440
	AATCCGTCA GATCCCTATG CTGACCACCT CCTTCAGAT CAACAGCTGC CCTTACTCGA	1500
50	GCTCAAGCTT CGAATTCTGC AGTCGACGGT ACCGCGGCCG CTAACATAATA GCCCATTCTC	1560
	CAAGGTACGT AGCGGGGATC AATTCCGCC CCCCCCTAAC GTTACTGGCC GAAGCCGCTT	1620
	GGAATAAGGC CGGTGTGCGT TTGTCTATAT GTTATTTCC ACCATATTGC CGTCTTTGG	1680
55	CAATGTGAGG GCCCGGAAAC CTGGCCCTGT CTTCTTGACG AGCATTCTA GGGGTCTTTC	1740
	CCCTCTCGCC AAAGGAATGC AAGGTCTGTT GAATGTGCGT AAGGAAGCAG TTCCCTCTGGA	1800
60	AGCTTCTTGA AGACAAACAA CGTCTGTAGC GACCCTTGCA AGGCAGCGGA ACCCCCCCACC	1860
	TGGCGACAGG TGCCTCTGCG GCCAAAAGCC ACGTGTATAA GATACACCTG CAAAGGGCGC	1920
	ACAACCCAG TGCCACGTTG TGAGTTGGAT AGTTGTGGAA AGAGTCAAAT GGCTCTCCTC	1980
65	AAGCGTATTG AACAAAGGGGC TGAAGGATGC CCAGAAGGTA CCCATTGTA TGGGATCTGA	2040

	TCTGGGCCT CGGTGCACAT GCTTTACATG TGTTTAGTCG AGGTTAAAAA AACGTCTAGG	2100
	CCCCCGAAC CACGGGGACG TGGTTTCCT TTGAAAACA CGATAACGGGA TCCACCGGTC	2160
5	GCCACCATGG GTAAAGGAGA AGAACCTTTC ACAGGAGTGTG TCCCATTCT TGTTGAATTA	2220
	GATGGTGATG TTAATGGCA CAAATTTCT GTCACTGGAG AGGGTGAAGG TGATGCAACA	2280
10	TACGGAAAAC TTACCCCTAA ATTTATTTCG ACTACTGGAA AACTACCTGT TCCATGGCCA	2340
	ACACTTGTCA CTACTTTCAC TTATGGTGT CAATGCTTT CAAGATAACCC AGATCATATG	2400
	AAACGGCATG ACTTTTCAA GAGTGCCATG CCCGAAGGTT ATGTACAGGA AAGAACTATA	2460
15	TTTTCAAAG ATGACGGAA CTACAAGACA CGTGCTGAAG TCAAGTTGA AGGTGATACC	2520
	CTTGTAAATA GAATCGAGTT AAAAGGTATT GATTTAAAG AAGATGGAAA CATTCTGGA	2580
20	CACAAATTGG AATACAAC TAACACACAC AATGTATACA TCATGGCAGA CAAACAAAAG	2640
	AATGGAACCA AAGTTAACTT CAAAATTAGA CACAACATTG AAGATGGAAG CGTTCAACTA	2700
	GCAGACCATT ATCAACAAAA TACTCCAATT GGCGATGGCC CTGTCCTTT ACCAGACAAAC	2760
25	CATTACCTGT CCACACAATC TGCCCTTCG AAAGATCCCA ACGAAAAGAG AGACCACATG	2820
	GTCCTCTTG AGTTGTAAAC AGCTGCTGGG ATTACACATG GCATGGATGA ACTATACAAG	2880
	TCCGGATCTA GATAACTGTA TCGATGGATC CGAAGGCAGG GACAGCAGTG CAGTGGTGA	2940
30	CAGAAAGCAA GTGATCTAGG CCAGCAGCCT CCCTAAAGGG ACTTCAGCCC ACAAAGCCAA	3000
	ACTTGTGGCT TTAATACAAG CTCTGTAAAT GGTAAAAAAA AAAAAGTCTA CACGGACAGC	3060
35	AGGTATGCTC TTGCCACTGT ACAGAGCAAT ATACAGACAA AGAGAACTGT TGACATCTGC	3120
	AGAGAAAGAC CTAAGATGCT GTGGCTAAA GAAATCAGAT GGCAAATCTA ACCGCCAGG	3180
	CATCCTAAAG AGCAATGATC CTGACAGTCT GAAGACTATC AAGTTATAGA CAAATTAAGA	3240
40	CTGGTAAAAAA AAACCTGTA TAAAATAGTA AAAACTGAAA AAAGAAAAGT AGTCCTCTCA	3300
	TGAGAAAGACA GACCTGACAT CTACTGAAAA ATAGACTTTA CTGGAAAAAA TATGTGTATG	3360
45	AATACCTTCT AGTTTTGTG AACGTTCTCA AGATGGATAA AAGCTTTCC TTGTAAAACG	3420
	AGACTGATCA GATAGTCATC AAGAAGATTG TTAAAGAAAA TTTCCAAGG TTCGGAGTGC	3480
	CAAAAGCAAT AGTGTCAAGAT AATGGCCTG CCTTGTTGC CCAGGTAAGT CAGGGTGTGG	3540
50	CCAAGTATT AGAGGTCAAA TGAAAATTCC ATTGTGTGTA CAGACCTCAG AGCTCAGGAA	3600
	AGATAAAAAAA GAATAAATAA AACTCTAAC AGACCTTGAC AAAATTAATC CTAGAGACTG	3660
55	GCACAGACTT ACTTGGTACT CCTTCCCTT GCCCTATTAA GAACTGAGAA TACTCCCTCT	3720
	TGATTCGGTT TTACTCTTT TAAGATCCTT TATGGGGCTC CTATGCCATC ACTGTCTAA	3780
	ATGATGTGTT TAAACCTATG TTGTTATAAT AATGATCTAT ATGTTAAGTT AAAAGGTTG	3840
60	CAGGTGGTGC AGAAAGAAGT CTGGTCACAA CTGGCTACAG TGAACAAGCT GGGTACCCCA	3900
	AGGACATCTT ACCAGTTCCA GCCAGAGATC TGATCTACGA TCCCCGGGTC GACCCGGGTC	3960
65	GACCCTGTGG AATGTGTGTC AGTTAGGGTG TGGAAAGTCC CCAGGCTCCC CAGCAGGCAG	4020
	AACTATGCAA AGCATGCATC TCAATTAGTC AGCAACCAGG TGTGGAAAGT CCCCAGGCTC	4080

	CCCAGCAGGC AGAAGTATGC AAAGCATGCA TCTCAATTAG TCAGCAACCA TAGTCCCGCC	4140
5	CCTAACTCCG CCCATCCCGC CCCTAACTCC GCCCAGTTCC GCCCATTCTC CGCCCCATGG	4200
	CTGACTAATT TTTTTTATTAT ATGCAGAGGC CGAGGCCGCC TCGGCCTCTG AGCTATTCCA	4260
	GAAGTAGTGA GGAGGCTTT TTGGAGGCCT AGGCTTTGC AAAAAGCTTC ACGCTGCCGC	4320
10	AAGCACTCAG GCGCAAGGG CTGCTAAAGG AAGCGGAACA CGTAGAAAGC CAGTCCGCAG	4380
	AAACGGTGCT GACCCCGGAT GAATGTCAGC TACTGGGCTA TCTGGACAAG GGAAAACGCA	4440
15	AGCGCAAAGA GAAAGCAGGT AGCTTGCAGT GGGCTTACAT GGCGATAGCT AGACTGGGCG	4500
	GTTTTATGGA CAGCAAGCGA ACCGGAATTG CCAGCTGGGG CGCCCTCTGG TAAGGTTGGG	4560
	AAGCCCTGCA AAGTAAACTG GATGGCTTTC TTGCCGCCAA GGATCTGATG GCGCAGGGGA	4620
20	TCAAGATCTG ATCAAGAGAC AGGATGAGGA TCGTTTCGCA TGATTGAACA AGATGGATTG	4680
	CACGCAGGTT CTCCGGCCGC TTGGGTGGAG AGGCTATTG GCTATGACTG GGCACAAACAG	4740
25	ACAATCGGCT GCTCTGATGC CGCCGTGTT CGGCTGTCAG CGCAGGGCG CCCGGTTCTT	4800
	TTTGTCAAGA CCGACCTGTC CGGTGCCCTG AATGAACTGC AGGACGAGGC AGCGCGGCTA	4860
	TCGTGGCTGG CCACGACGGG CGTTCCCTGTC GCAGCTGTGC TCGACGTTGT CACTGAAGCG	4920
30	GGAAGGGACT GGCTGCTATT GGGCGAAGTG CCGGGGCAGG ATCTCCTGTC ATCTCACCTT	4980
	GCTCCTGCCG AGAAAGTATC CATCATGGCT GATGCAATGC GGCGGCTGCA TACGCTTGAT	5040
	CCGGCTACCT GCCCATTGCA CCACCAAGCG AAACATCGCA TCGAGCGAGC ACGTACTCGG	5100
35	ATGGAAGCCG GTCTTGTGCA TCAGGATGAT CTGGACGAAG AGCATCAGGG GCTCGGCCA	5160
	GCCGAACGT TCGCCAGGCT CAAGGCGCGC ATGCCGACG GCGAGGATCT CGTCGTGACC	5220
40	CATGGCGATG CCTGCTTGCC GAATATCATG GTGGAAAATG GCCGCTTTTC TGGATTCATC	5280
	GACTGTGGCC GGCTGGGTGT GGC GGACCGC TATCAGGACA TAGCGTTGGC TACCCGTGAT	5340
	ATTGCTGAAG AGCTTGGCGG CGAATGGCT GACCGCTTCC TCGTGCTTTA CGGTATCGCC	5400
45	GCTCCCGATT CGCAGCGCAT CGCCTTCTAT CGCCTTCTTG ACGAGTTCTT CTGAGCGGGA	5460
	CTCTGGGGTT CGAAATGACC GACCAAGCGA CGCCCAACCT GCCATCACGA GATTCGATT	5520
50	CCACCGCCGC CTTCTATGAA AGGTTGGCT TCGGAATCGT TTTCCGGAC GGAATTGTA	5580
	ATCTGCTGCT TGCAAACAAA AAAACCACCG CTACCAGCGG TGGTTGTTT GCCGGATCAA	5640
	GAGCTACCAA CTCTTTTCC GAAGGTAACG GGCTTCAGCA GAGCGCAGAT ACCAAATACT	5700
55	GTCCTTCTAG TGTAGCCGTA GTTAGGCCAC CACTTCAAGA ACTCTGTAGC ACCGCCTACA	5760
	TACCTCGCTC TGCTAACCT GTTACCAAGTG GCTGCTGCCA GTGGCGATAA GTCGTGTCTT	5820
60	ACCGGGTTGG ACTCAAGACG ATAGTTACCG GATAAGGCAGC AGCGGTCGGG CTGAACGGGG	5880
	GGTTCGTGCA CACAGCCCAG CTTGGAGCGA ACGACCTACA CCGAACTGAG ATACCTACAG	5940
	CGTGAGCATT GAGAAAGCGC CACGCTTCCC GAAGGGAGAA AGGCGGACAG GTATCCGGTA	6000
65	AGCGGCAGGG TCGGAACAGG AGAGCGCACG AGGGAGCTTC CAGGGGGAAA CGCCTGGTAT	6060

CTTTATAAGTC CTGTCGGGTT TCGCCACCTC TGACTTGAGC GTCGATTTTT GTGATGCTCG 6120  
 TCAGGGGGGC GGAGCCTATG GAAAAACGCC AGCAACGCCG AGATGCGCCG CCTCGAGAAC 6180  
 5 CCTGGCCCTA TTATTGGGTG GACTAACCAT GGGGGGAATT GCCGCTGGAA TAGGAACAGG 6240  
 GACTACTGCT CTAATGGCCA CTCAGCAATT CCAGCAGCTC CAAGCCGCAG TACAGGATGA 6300  
 10 TCTCAGGGAG GTTGAAAAAT CAATCTCTAA CCTAGAAAAG TCTCTCACTT CCCTGTCTGA 6360  
 AGTTGTCTA CAGAATCGAA GGGGCCTAGA CTTGTTATT CTAAAAGAAC GAGGGCTGTG 6420  
 TGCTGCTCTA AAAGAAGAAC GTTGCTTCTA TGCGGACCAC ACAGGACTAG TGAGAGACAG 6480  
 15 CATGGCCAAA TTGAGAGAGA GGCTTAATCA GAGACAGAAA CTGTTGAGT CAACTCAAGG 6540  
 ATGGTTGAG GGACTGTTA ACAGATCCCC TTGGTTTACC ACCTTGATAT CTACCATTAT 6600  
 20 GGGACCCCTC ATTGTACTCC TAATGATTTC GCTCTTCGGA CCCTGCATTC TTAATCGATT 6660  
 AGTCCAATT GTTAAAGACA GGATATCAGT GGTCCAGGCT CTAGTTTGA CTCAACAATA 6720  
 TCACCAGCTG AAGCCTATAG AGTACGAGCC ATAGATAAAA TAAAAGATT TATTTAGTCT 6780  
 25 CCAGAAAAAG GGGGG 6795  
 (2) INFORMATION FOR SEQ ID NO:23:  
 (i) SEQUENCE CHARACTERISTICS:  
 30 (A) LENGTH: 9093 base pairs  
 (B) TYPE: nucleic acid  
 (C) STRANDEDNESS: single  
 (D) TOPOLOGY: linear  
 (ii) MOLECULE TYPE: DNA (genomic)  
 35

## (xi) SEQUENCE DESCRIPTION: SEQ ID NO:23:

40 AATGAAAGAC CCCACCTGTA GGTTGGCAA GCTAGCTTAA GTAACGCCAT TTTGCAAGGC 60  
 ATGGAAAAAT ACATAACTGA GAATAGAGAA GTTCAGATCA AGGTCAGGAA CAGATGGAAC 120  
 45 AGCTGAATAT GGGCAAACA GGATATCTGT GGTAAGCAGT TCCTGCCCG GCTCAGGGCC 180  
 AAGAACAGAT GGAACAGCTG AATATGGCC AAACAGGATA TCTGTGGTAA GCAGTTCCCTG 240  
 50 CCCCGGCTCA GGGCCAAGAA CAGATGGTCC CCAGATGCCG TCCAGCCCTC AGCAGTTCT 300  
 AGAGAACCAT CAGATGTTTC CAGGGTGCCC CAAGGACCTG AAATGACCCCT GTGCCTTATT 360  
 TGAACTAACC AATCAGTTCG CTTCTCGCTT CTGTTCGCGC GCTCTGCTC CCCGAGCTCA 420  
 55 ATAAAAGAGC CCACAACCCC TCACTCGGGG CGCCAGTCCT CCGATTGACT GAGTCGCCCG 480  
 GGTACCCGTG TATCCAATAA ACCCTCTTGC AGTTGCATCC GACTTGTGGT CTCGCTGTT 540  
 60 CTTGGGAGGG TCTCCTCTGA GTGATTGACT ACCCGTCAGC GGGGGTCTTT CATTGGGGG 600  
 CTCGTCCGGG ATCGGGAGAC CCCTGCCAG GGACCACCGA CCCACCACCG GGAGGTAAGC 660  
 TGGCCAGCAA CTTATCTGTG TCTGTCCGAT TGTCTAGTGT CTATGACTGA TTTTATGCGC 720  
 65 CTGCGTCGGT ACTAGTTAGC TAACTAGCTC TGTATCTGGC GGACCCGTGG TGGAACGTAC 780

	GAGTTGGAA CACCCGGCCG CAACCCTGGG AGACGTCCA GGAGGAACAG GGGAGGATCA	840
	GGGACGCCTG GTGGACCCCT TTGAAGGCCA AGAGACCATT TGGGGTTGCG AGATCGTGGG	900
5	TTCGAGTCCC ACCTCGTGCC CAGTTGCGAG ATCGTGGGTT CGAGTCCCAC CTCGTGTTT	960
	GTTGCGAGAT CGTGGGTTCG AGTCCCACCT CGCGTCTGGT CACGGGATCG TGGGTTGAG	1020
10	TCCCACCTCG TGTTTGTG CGAGATCGTG GGTTGAGTC CCACCTCGCG TCTGGTCACG	1080
	GGATCGTGGG TTCGAGTCCC ACCTCGTGCA GAGGGTCTCA ATTGGCCGGC CTTAGAGAGG	1140
	CCATCTGATT CTTCTGGTTT CTCTTTTGT CTTAGTCTCG TGTCCGCTCT TGTTGTGACT	1200
15	ACTGTTTTC TAAAAATGGG ACAATCTGTG TCCACTCCCC TTTCTCTGAC TCTGGTTCTG	1260
	TCGCTTGGTA ATTTTGTGG TTTACGTTG TTTTGTGAG TCGTCTATGT TGTCTGTTAC	1320
20	TATCTTGTGTT TTGTTGTGG TTTACGGTTT CTGTGTGTGT CTTGTGTGTC TCTTGTGTT	1380
	CAGACTTGGG CTGATGACTG ACGACTGTTT TTAAGTTATG CCTTCTAAAA TAAGCCTAAA	1440
	AATCCTGTCA GATCCCTATG CTGACCACTT CCTTCAGAT CAACAGCTGC CCTTACGTAT	1500
25	CGATGGATCC CTCGACTAAC TAATAGCCA TTCTCCAAGG TCGAGCGGGG TCAATTCCGC	1560
	CCCCCCCTA ACGTTACTGG CCGAAGCCGC TTGGAATAAG GCCGGTGTGC GTTGTCTAT	1620
30	ATGTTATTTT CCACCATATT GCCGTCTTTT GGCAATGTGA GGGCCCGGAA ACCTGGCCCT	1680
	GTCTTCTTGA CGAGCATTCC TAGGGGTCTT TCCCCTCTCG CCAAAGGAAT GCAAGGTCTG	1740
	TTGAATGTGCG TGAAGGAAGC AGTCCTCTG GAAGCTTCTT GAAGACAAAC AACGTCTGTA	1800
35	GCGACCCTTT GCAGGCAGCG GAACCCCCCA CCTGGCGACA GGTGCCTCTG CGGCCAAAAG	1860
	CCACGTGTAT AAGATACACC TGCAAAGGCG GCACAACCCC AGTGCCACGT TGTGAGTTGG	1920
40	ATAGTTGTGG AAAGAGTCAA ATGGCTCTCC TCAAGCGTAT TCAACAAGGG GCTGAAGGAT	1980
	GCCCAGAAGG TACCCATTG TATGGATCT GATCTGGGGC CTCGGTGCAC ATGCTTTACA	2040
	TGTGTTAGT CGAGGTTAAA AAAACGTCTA GGCCCCCGA ACCACGGGGG CGTGGTTTC	2100
45	CTTTGAAAAA CACGATAATA ATCATGGCG CGGATCCCGT CGTTTACAA CGTCGTGACT	2160
	GGGAAAACCC TGGCGTTACC CAACTTAATC GCCTTGCAGC ACATCCCCCT TTCGCCAGCT	2220
50	GGCGTAATAG CGAAGAGGCC CGCACCGATC GCCCTTCCA ACAGTTGCGC AGCCTGAATG	2280
	GCGAATGGCG CTTTGCCTGG TTTCCGGCAC CAGAAGCGGT GCCGGAAAGC TGGCTGGAGT	2340
	GCGATCTTCC TGAGGCCGAT ACTGTCGTG TCCCCTCAAA CTGGCAGATG CACGGTTACG	2400
55	ATGCGCCCAT CTACACCAAC GTAACCTATC CCATTACGGT CAATCCGCCG TTTGTTCCCA	2460
	CGGAGAATCC GACGGGTTGT TACTCGCTCA CATTAAATGT TGATGAAAGC TGGCTACAGG	2520
60	AAGGCCAGAC GCGAATTATT TTTGATGGCG TTAACTCGGC GTTTCATCTG TGGTGCAACG	2580
	GGCGCTGGGT CGGTTACGGC CAGGACAGTC GTTTCGGTC TGAATTGAC CTGAGCGCAT	2640
	TTTTACGCGC CGGAGAAAAC CGCCTCGCGG TGATGGTGCT GCGTTGGAGT GACGGCAGTT	2700
65	ATCTGGAAGA TCAGGATATG TGGCGGATGA CGGGCATTTT CCGTGACGTC TCGTTGCTGC	2760

	ATAAACCGAC TACACAAATC AGCGATTCC ATGTTGCCAC TCGCTTTAAT GATGATTCA	2820
	GCCGCGCTGT ACTGGAGGCT GAAGTTCAGA TGTGCGGCCGA GTTGCCTGAC TACCTACGGG	2880
5	TAACAGTTTC TTTATGGCAG GGTGAAACGC AGGTGCCAG CGGCACCGCG CCTTTCGGCG	2940
	GTGAAATTAT CGATGAGCGT GGTGGTTATG CCGATCGCGT CACACTACGT CTGAACGTCG	3000
10	AAAACCGAA ACTGTGGAGC GCCGAAATCC CGAACATCTCA TCGTGCCTGT GTTGAACGTGC	3060
	ACACCGCCGA CGGCACGCTG ATTGAAGCAG AAGCCTGCGA TGTCGGTTTC CGCGAGGTGC	3120
	GGATTGAAAA TGGTCTGCTG CTGCTGAACG GCAAGCCGTT GCTGATTGGA GGCGTTAACCG	3180
15	GTCACGAGCA TCATCCTCTG CATGGTCAGG TCATGGATGA GCAGACGATG GTGCAGGATA	3240
	TCCTGCTGAT GAAGCAGAAC AACTTTAACG CCGTGCCTG TTGCGATTAT CCGAACCATC	3300
20	CGCTGTGGTA CACGCTGTGC GACCCTACG GCCTGTATGT GGTGGATGAA GCCAATATTG	3360
	AAACCCACGG CATGGTGCCA ATGAATCGTC TGACCGATGA TCCGCGCTGG CTACCGCGA	3420
	TGAGCGAACG CGTAACCGCA ATGGTGCAGC GCGATCGTA TCACCCGAGT GTGATCATCT	3480
	GGTCGCTGGG GAATGAATCA GGCCACGGCG CTAATCACGA CGCGCTGTAT CGCTGGATCA	3540
25	AATCTGTCGA TCCTTCCCAGT CCGGTGCAGT ATGAAGGCCG CGGAGCCGAC ACCACGGCCA	3600
	CCGATATTAT TTGCCCCATG TACGCGCGCG TGGATGAAGA CCAGCCCTTC CCGGCTGTGC	3660
30	CGAAATGGTC CATCAAAAAA TGGCTTCGC TACCTGGAGA GACGCGCCCG CTGATCCTT	3720
	CGCAATACGC CCACCGGATG GGTAACAGTC TTGGCGGTT CGCTAAATAC TGGCAGGCCT	3780
35	TTCGTCAGTA TCCCCGTTA CAGGGCGGCT TCGTCTGGGA CTGGGTGGAT CAGTCGCTGA	3840
	TTAAATATGA TGAAAACGGC AACCCGTGGT CGGCTTACGG CGGTGATTT GGCGATACGC	3900
	CGAACGATCG CCAGTTCTGT ATGAACGGTC TGGTCTTGC CGACCGCACG CCGCATCCAG	3960
40	CGCTGACGGA AGCAAAACAC CAGCAGCAGT TTTTCCAGTT CCGTTTATCC GGGCAAACCA	4020
	TCGAAGTGAC CAGCGAATAC CTGTTCCGTC ATAGCGATAA CGAGCTCCTG CACTGGATGG	4080
45	TGGCGCTGGA TGGTAAGCCG CTGGCAAGCG GTGAAGTGCC TCTGGATGTC GCTCCACAAG	4140
	GTAAACAGTT GATTGAACGT CCTGAACATAC CGCAGCCGGA GAGCGCCGGG CAACTCTGGC	4200
	TCACAGTACG CGTAGTGCAA CGAACGCGA CCGCATGGTC AGAAGCCGGG CACATCAGCG	4260
50	CCTGGCAGCA GTGGCGTCTG GCGGAAAACC TCAGTGTGAC GCTCCCCGCC GCGTCCCACG	4320
	CCATCCCGCA TCTGACCACC AGCGAAATGG ATTTTGAT CGAGCTGGGT AATAAGCGTT	4380
55	GGCAATTAA CCGCCAGTCA GGCTTCTTT CACAGATGTG GATTGGCGAT AAAAAACAAC	4440
	TGCTGACGCC GCTGCGCGAT CAGTCACCC GTGCACCGCT GGATAACGAC ATTGGCGTAA	4500
	GTGAAGCGAC CCGCATTGAC CCTAACGCCT GGGTCGAACG CTGGAAGGCG GCGGGCCATT	4560
60	ACCAGGCCGA AGCAGCGTTG TTGCAGTGCA CGGCAGATAC ACTTGCTGAT GCGGTGCTGA	4620
	TTACGACCGC TCACCGCTGG CAGCATCAGG GGAAAACCTT ATTTATCAGC CGGAAAACCT	4680
65	ACCGGATTGA TGGTAGTGGT CAAATGGCGA TTACCGTTGA TGTGAAAGTG GCGAGCGATA	4740
	CACCGCATCC GGCGCGGATT GGCCTGAACG GCCAGCTGGC GCAGGTAGCA GAGCGGGTAA	4800

	ACTGGCTCGG ATTAGGGCCG CAAGAAAACT ATCCCGACCG CCTTACTGCC GCCTGTTTG	4860
5	ACCGCTGGGA TCTGCCATTG TCAGACATGT ATACCCCGTA CGTCTTCCCG AGCGAAAACG	4920
	GTCTGCCTG CGGGACGCGC GAATTGAATT ATGGCCCACA CCAGTGGCGC GGCGACTTCC	4980
	AGTTCAACAT CAGCCGCTAC AGTCAACAGC AACTGATGGA AACCAAGCCAT CGCCATCTGC	5040
10	TGCACGCGGA AGAAGGCACA TGGCTGAATA TCGACGGTTT CCATATGGGG ATTGGTGGCG	5100
	ACGACTCCTG GAGCCCGTCA GTATCGGCGG AATTCCAGCT GAGCGCCGGT CGCTACCATT	5160
15	ACCAGTTGGT CTGGTGTCAA AAATAATAAT AACCGGGCAG GGGGGATCCG AAGGCGGGGA	5220
	CAGCAGTGCA GTGGTGGACA GAAAGCAAGT GATCTAGGCC AGCAGCCTCC CTAAAGGGAC	5280
	TTCAGCCCAC AAAGCCAAAC TTGTGGCTTT AATACAAGCT CTGTAATGG TAAAAAAA	5340
20	AAAGTCTACA CGGACAGCAG GTATGCTCTT GCCACTGTAC AGAGCAATAT ACAGACAAAG	5400
	AGAACTGTTG ACATCTGCAG AGAAAGACCT AAGATGCTGT GGCTAAAAGA AATCAGATGG	5460
	CAAATCTAAC CGCCCAGGCA TCCTAAAGAG CAATGATCCT GACAGTCTGA AGACTATCAA	5520
25	GTTATAGACA AATTAAGACT GGTAAAAAAA ACCCTGTATA AAATAGTAAA AACTGAAAAA	5580
	AGAAAACCTAG TCCTCTCATG AGAAGACAGA CCTGACATCT ACTGAAAAAT AGACTTTACT	5640
	GGAAAAAAATA TGTGTATGAA TACCTTCTAG TTTTGTTGAA CGTTCTCAAG ATGGATAAAA	5700
30	GCTTTCCCTT GTAAAACGAG ACTGATCAGA TAGTCATCAA GAAGATTGTT AAAGAAAATT	5760
	TTCCAAGGTT CGGAGTGCCA AAAGCAATAG TGTCAGATAA TGGTCCTGCC TTTGTTGCC	5820
35	AGGTAAGTCA GGGTGTGGCC AAGTATTTAG AGGTCAAATG AAAATTCCAT TGTGTGTACA	5880
	GACCTCAGAG CTCAGGAAAG ATAAAAAAAGA ATAAATAAAA CTCTAACAG ACCTTGACAA	5940
	AATTAATCCT AGAGACTGGC ACAGACTTAC TTGGTACTCC TTCCCCTTGC CCTATTTAGA	6000
40	ACTGAGAATA CTCCCTCTTG ATTGGTTTT ACTCTTTTA AGATCCTTTA TGGGCTCCT	6060
	ATGCCATCAC TGTCTTAAAT GATGTGTTA AACCTATGTT GTTATAATAA TGATCTATAT	6120
45	GTAAAGTTAA AAGGCTTGCA GGTGGTCAG AAAGAAGTCT GGTACAACG GGCTACAGTG	6180
	AACAAGCTGG GTACCCCAAG GACATCTTAC CAGTTCCAGC CAGAGATCTG ATCTACGATC	6240
	CCCGGGTCGA CCCGGGTCGA CCCTGTGGAA TGTGTGTCAG TTAGGGTGTG GAAAGTCCCC	6300
50	AGGCTCCCCA GCAGGCAGAA GTATGCAAAG CATGCATCTC AATTAGTCAG CAACCAGGTG	6360
	TGGAAAGTCC CCAGGGCTCCC CAGCAGGCAG AAGTATGCCA AGCATGCATC TCAATTAGTC	6420
55	AGCAACCATA GTCCCCCCCC TAACTCCGCC CATCCGCC CTAACTCCGC CCAGTCCGC	6480
	CCATTCTCCG CCCCATGGCT GACTAATTTC TTTTATTAT GCAGAGGCCG AGGCCGCC	6540
	GGCCTCTGAG CTATTCCAGA AGTAGTGAGG AGGCTTTTT GGAGGCCTAG GCTTTGCAA	6600
60	AAAGCTTCAC GCTGCCGCAA GCACTCAGGG CGCAAGGGCT GCTAAAGGAA GCGGAACACG	6660
	TAGAAAGCCA GTCCGCAGAA ACGGTGCTGA CCCCGGATGA ATGTCAGCTA CTGGGCTATC	6720
65	TGGACAAGGG AAAACGCAAG CGCAAAGAGA AAGCAGGTAG CTTGCAGTGG GCTTACATGG	6780

	CGATAGCTAG ACTGGGCGGT TTTATGGACA GCAAGCGAAC CGGAATTGCC AGCTGGGCG	6840
	CCCTCTGGTA AGGTTGGAA GCCCTGCAA GTAAACTGGA TGGCTTCTT GCCGCCAAGG	6900
5	ATCTGATGGC GCAGGGGATC AAGATCTGAT CAAGAGACAG GATGAGGATC GTTCGCATG	6960
	ATTGAACAAG ATGGATTGCA CGCAGGTTCT CCGGCCGCTT GGGTGGAGAG GCTATTGGC	7020
10	TATGACTGGG CACAACAGAC AATCGGCTGC TCTGATGCCG CCGTGTCCG GCTGTCAGCG	7080
	CAGGGGCGCC CGGTTCTTT TGTCAAGACC GACCTGTCCG GTGCCCTGAA TGAACTGCAG	7140
	GACGAGGCAG CGCGGCTATC GTGGCTGGCC ACGACGGGCG TTCCCTGCGC AGCTGTGCTC	7200
15	GACGTTGTCA CTGAAGCGGG AAGGGACTGG CTGCTATTGG GCGAAGTGCC GGGCAGGAT	7260
	CTCCTGTCAT CTCACCTTGC TCCTGCCGAG AAAGTATCCA TCATGGCTGA TGCAATGCGG	7320
20	CGGCTGCATA CGCTTGATCC GGCTACCTGC CCATTGACCC ACCAAGCGAA ACATCGCATC	7380
	GAGCGAGCAC GTACTCGGAT GGAAGCCGGT CTTGTCGATC AGGATGATCT GGACGAAGAG	7440
	CATCAGGGGC TCGGCCAGC CGAACTGTTG GCCAGGCTCA AGGCGCGCAT GCCCCGACGGC	7500
25	GAGGATCTCG TCGTGACCCA TGGCGATGCC TGCTTGCCGA ATATCATGGT GGAAAATGGC	7560
	CGCTTTCTG GATTGATCGA CTGTGGCCGG CTGGGTGTGG CGGACCGCTA TCAGGACATA	7620
	GCGTTGGCTA CCCGTGATAT TGCTGAAGAG CTTGGCGGCG AATGGGCTGA CCGCTTCCTC	7680
30	GTGCTTACG GTATGCCGC TCCCGATTG CAGCGCATCG CCTTCTATCG CCTTCTTGAC	7740
	GAGTTCTTCT GAGCGGGACT CTGGGGTTCG AAATGACCGA CCAAGCGACG CCCAACCTGC	7800
35	CATCACGAGA TTTCGATTCC ACCGCCGCCT TCTATGAAAG GTTGGGCTTC GGAATCGTT	7860
	TCCGGGACGG AATTGTAAT CTGCTGCTTG CAAACAAAAA AACCAACCGCT ACCAGCGGTG	7920
	GTGGTTGC CGGATCAAGA GCTACCAACT CTTTTCCGA AGGTAACTGG CTTCAGCAGA	7980
40	GCGCAGATAAC CAAATACTGT CCTTCTAGTG TAGCCGTAGT TAGGCCACCA CTTCAAGAAC	8040
	TCTGTAGCAC CGCCTACATA CCTCGCTCTG CTAATCCTGT TACCAAGTGGC TGCTGCCAGT	8100
45	GGCGATAAGT CGTGTCTTAC CGGGTTGGAC TCAAGACGAT AGTTACCGGA TAAGGCGCAG	8160
	CGGTCGGGCT GAACGGGGGG TTCGTGCACA CAGCCCAGCT TGGAGCGAAC GACCTACACC	8220
	GAACTGAGAT ACCTACAGCG TGAGCATTGA GAAAGCGCCA CGCTTCCGA AGGGAGAAAG	8280
50	GCGGACAGGT ATCCGGTAAG CGGCAGGGTC GGAACAGGAG AGCGCACGAG GGAGCTTCCA	8340
	GGGGGAAACG CCTGGTATCT TTATAGTCCT GTCGGGTTTC GCCACCTCTG ACTTGAGCGT	8400
55	CGATTTTGT GATGCTCGTC AGGGGGCGG AGCCTATGGA AAAACGCCAG CAACGCCAG	8460
	ATGCGCCGCC TCGAGAACCC TGGCCCTATT ATTGGGTGGA CTAACCATGG GGGGAATTGC	8520
	CGCTGGAATA GGAACACAGGA CTACTGCTCT AATGCCACT CAGCAATTCC AGCAGCTCCA	8580
60	AGCCGCAGTA CAGGATGATC TCAGGGAGGT TGAAAATCA ATCTCTAACC TAGAAAAGTC	8640
	TCTCACTTCC CTGTCTGAAG TTGTCTTACA GAATCGAAGG GGCCTAGACT TGTTATTCT	8700
65	AAAAGAAGGA GGGCTGTGTG CTGCTCTAAA AGAAGAATGT TGCTTCTATG CGGACCACAC	8760
	AGGACTAGTG AGAGACAGCA TGGCCAAATT GAGAGAGAGG CTTAATCAGA GACAGAAACT	8820

5            GTTTGAGTCA ACTCAAGGAT GGTTTGAGGG ACTGTTAAC AGATCCCTT GGTTTACAC        8880  
           CTTGATATCT ACCATTATGG GACCCCTCAT TGTACTCCTA ATGATTTGC TCTTCGGACC        8940  
           CTGCATTCTT AATCGATTAG TCCAATTGT TAAAGACAGG ATATCAGTGG TCCAGGCTCT        9000  
           AGTTTGACT CAACAATATC ACCAGCTGAA GCCTATAGAG TACGAGCCAT AGATAAAATA        9060  
 10          AAAGATTTA TTTAGTCTCC AGAAAAAGGG GGG                                        9093

## (2) INFORMATION FOR SEQ ID NO:24:

15          (i) SEQUENCE CHARACTERISTICS:  
             (A) LENGTH: 46 base pairs  
             (B) TYPE: nucleic acid  
             (C) STRANDEDNESS: single  
             (D) TOPOLOGY: linear

20          (ii) MOLECULE TYPE: DNA (genomic)

## (xi) SEQUENCE DESCRIPTION: SEQ ID NO:24:

25          GACTAACCTT GATTCCCTGG AGGCGGGGGT CTTTCATTG GGGGCT                        46

## (2) INFORMATION FOR SEQ ID NO:25:

30          (i) SEQUENCE CHARACTERISTICS:  
             (A) LENGTH: 4834 base pairs  
             (B) TYPE: nucleic acid  
             (C) STRANDEDNESS: single  
             (D) TOPOLOGY: linear

35          (ii) MOLECULE TYPE: DNA (genomic)

## (xi) SEQUENCE DESCRIPTION: SEQ ID NO:25:

40          TGAAGAATAA AAAATTACTG GCCTCTTGTG AGAACATGAA CTTTCACCTC GGAGCCCACC        60  
 45          CCCTCCCATC TGGAAAACAT ACTTGAGAAA AACATTTCT GGAACAAACCA CAGAATGTTT        120  
           CAACAGGCCA GATGTATTGC CAAACACAGG ATATGACTCT TTGGTTGAGT AAATTGTGG        180  
           TTGTTAAACT TCCCCTATT CCTCCCCATT CCCCTCCCA GTTGTGGTT TTTCCCTTA        240  
           AAAGCTTGTG AAAAATTGA GTCGTCGTCG AGACTCCTCT ACCCTGTGCA AAGGTGTATG        300  
           AGTTTCGACC CCAGAGCTCT GTGTGCTTTC TGTTGCTGCT TTATTCGAC CCCAGAGCTC        360  
 55          TGGTCTGTGT GCTTTCATGT CGCTGCTTTA TTAAATCTTA CCTTCTACAT TTTATGTATG        420  
           GTCTCAGTGT CTTCTTGGGT ACGCGGCTGT CCCGGGACTT GAGTGTCTGA GTGAGGGTCT        480  
           TCCCTCGAGG GTCTTCATT TGGTACATGG GCCGGGAATT CGAGAATCTT TCATTTGGTG        540  
           CATTGGCCGG GAATTCGAAA ATCTTCATT TGGTGCATTG GCCGGGAAAC AGCGCGACCA        600  
           CCCAGAGGTC CTAGACCCAC TTAGAGGTAA GATTCTTGT TCTGTTTGG TCTGATGTCT        660  
 65          GTGTTCTGAT GTCTGTGTTG TGTTCTAAG TCTGGTGCAG TCGCAGTTTC AGTTTGCAG        720

	ACGCTCAGTG AGACCGCGCT CCGAGAGGGA GTGCCGGGTG GATAAGGATA GACGTGTCCA	780
	GGTGTCCACC GTCCGTTCGC CCTGGGAGAC GTCCCAGGAG GAACAGGGGA GGATCAGGGA	840
5	CGCCTGGTGG ACCCCTTGA AGGCCAAGAG ACCATTTGGG GTTGCAGAGAT CGTGGGTTCG	900
	AGTCCCACCT CGTCCCCAGT TGCGAGATCG TGGGTTCGAG TCCCACCTCG TGTTTGTG	960
10	CGAGATCGTG GTTTCGAGTC CCACCTCGCG TCTGGTCACG GGATCGTGGG TTCGAGTCCC	1020
	ACCTCGTGT TTGTTGCGAG ATCGTGGGTT CGAGTCCCAC CTCGCGTCTG GTCACGGGAT	1080
	CGTGGGTTCG AGTCCCACCT CGTGAGAGG GTCTCAATTG GCCGGCCTTA GAGAGGCCAT	1140
15	CTGATTCTTC TGGTTCTCT TTTGTCTTA GTCTCGTGT CGCTCTTGT GTGACTACTG	1200
	TTTTCTAAA AATGGGACAA TCTGTGTCCA CTCCCCTTTC TCTGACTCTG GTTCTGTCGC	1260
20	TTGGTAATTT TGTTGTTTA CGTTTGTGTT TGTGAGTCGT CTATGTTGTC TGTTACTATC	1320
	TTGTTTTGT TTGTTGTTA CGGTTCTGT GTGTGTCTTG TGTGTCTCTT TGTGTTCAAGA	1380
	CTTGGACTGA TGACTGACGA CTGTTTTAA GTTATGCCTT CTAAAATAAG CCTAAAATC	1440
25	CTGTCAGATC CCTATGCTGA CCACTTCCTT TCAGATCAAC AGCTGCCCTG CCTCCCACTC	1500
	CAAECTCCAGA GAGCAGCCAG CGGGTCACAG TGGTCCCGCC CATGAACCTG GAGCCTAGGG	1560
	AAAAATGAGC TCGGAAATCC GGAGCAAATG AGGAGTGGTC CCTGAGAAAGT CAGTGGCCTA	1620
30	AATGTTGTGG CTGCTGAAGC AAAAGAAGAG GAGGCTGTTG GAGTAGCCGG CCAAGAGCGC	1680
	CGCGGGTTCC CAGGCAGCTT CTCATCCCC TGTCCCTCCC ATCCCGTCTC TTGTTAACAG	1740
	AAAAACTGCT TTCACTTTGA GATATGAGTG GCCCGATACA GCCAGCTGT AGAGCTGTAC	1800
35	TCCCTTCCCT GCCCCACGTG TTTCTCTTC TCAGGCGACC CCTCCCTGAG CTGCTGGCAG	1860
	TGAGTCTGTT CTAAGCTCCA GTGAGGGAGG CATCCGCCA CTTGGGGCTT CTGTCCAAGG	1920
40	TAAGGAGCAC CTGTGAGTCT AACTGCCAGG CTCTGATGGG GGTCTCGTCT CTGTGGACT	1980
	AGAAAGTGTG CCAACAATCT GACCAAGGTA ACAGGAAGTT AAGACAAAGA CAGAGACCAA	2040
45	AGTCAGAACATC AGAGCTGTGC TGTGAGACAA AAAGATAAAA AAAATAAAAT GCTGGCCACA	2100
	AAAGTCAGGA AAACTAGAAA ACTTAGATAG TACCTGGCAA CAAAAGAAAG CTTTGGCTA	2160
	AAGATCAACG TGTATACTGT AAAGAAAATG AGCACTGGGT GAGAGACTGC CCCAACAAAA	2220
50	AGAAGAGGAG CCCCCCTCAT GACCAAACCC TTCACCTGTT CGTGGCTAAA AGTAAAGAGA	2280
	TAACAAAAGG GGTGCTAACCA CAGAAGCTGA GTCCTAAAA GAGTCCGGTG GCCTACCTGT	2340
	TGAAGCAGCT AAAAAAGAGA CTGTGTTCA TACTCCTCCA CTGACCAGTG CAAAACAAGC	2400
55	TAAAAAGTTC CTGGGCACTG CGGGCTTTG CAGATTGTGG ATTCCAGGTT TTGCTGAGTT	2460
	AAAGAGATAA ACAGCCCTTC GTATAGAAAA ATAAAAAACAA ACCTTGGATG TCCTTGGATG	2520
60	CTATTGAGAC TGCCCTAATG TTGTCCCCAG CTATGGGACT CCTAGATGTG ACTGAGAAC	2580
	AAGGTATTGC CAAAGAAGTT CTTACTCAGA GATTGGGACC CTGAAAAAGA CCTGTGGCAT	2640
65	ACTTGTAAAGA AATTAGACCT GGTGGCTGTA AGATGGCCTG CTTGTCTGCA CATAGTGGCT	2700
	TCTGGTCAAG GACGCAGATA AATTGACTCT GAGACAAAAC TTGGCACATG TCCTAGAAAG	2760

	TGTGGTTCAG	CCCCCATGAC	CGATGGCTGA	CTAACGCTCT	TGAAAACATT	ATCCAACTGT	2820
5	TCCCCTGACC	GATGGACACA	TTGTCAGAGC	TTTTTTGAC	TGAACGAGTG	ACCTTCGCTC	2880
	CCCCTGCTAT	CCTCGATCTC	ACTACTGCCT	GAGACTTCAC	CTACTCATCA	TTGTGCTGAC	2940
	ATTCTGGCAG	AAGAAAATCA	TACTCGAAAT	GATCTGAAGG	ATCAGATCAG	CCTTGGCCTG	3000
10	AGAGTTGAG	CTGGTACACG	GATGGCAGTA	GCCTGGAGGT	TAAGGGTAAG	CGGAAGGCCG	3060
	GGACAGCAGT	GCAGTGGTGG	ACAGAAAGCA	AGTGATCTAG	GCCAGCAGCC	TCCCTAAAGG	3120
15	GACTTCAGCC	CACAAAGCCA	AACTTGTGGC	TTAATACAA	GCTCTGTAAA	TGGTAAAAAA	3180
	AAAAAAAGTCT	ACACGGACAG	CAGGTATGCT	CTTGCCACTG	TACAGAGCAA	TATACAGACA	3240
	AAGAGAACTG	TTGACATCTG	CAGAGAAAGA	CCTAAGATGC	TGTGGCTAAA	AGAAATCAGA	3300
20	TGGCAAATCT	AACCGCCCAG	GCATCCTAAA	GAGCAATGAT	CCTGACAGTC	TGAAGACTAT	3360
	CAAGTTATAG	ACAAATTAAG	ACTGGTAAAAA	AAAACCTGT	ATAAAATAGT	AAAAACTGAA	3420
	AAAAGAAAAC	TAGTCCTCTC	ATGAGAAGAC	AGACCTGACA	TCTACTGAAA	AATAGACTTT	3480
25	ACTGGAAAAAA	ATATGTGTAT	GAATACCTTC	TAGTTTTGT	GAACGTTCTC	AAGATGGATA	3540
	AAAGCTTTTC	CTTGTAAAAC	GAGACTGATC	AGATAGTCAT	CAAGAAGATT	GTTAAAGAAA	3600
30	ATTTTCCAAG	GTTGGAGTG	CCAAAAGCAA	TAGTGTAGA	TAATGGCCT	GCCTTTGTTG	3660
	CCCAGGTAAG	TCAGGGTGTG	GCCAAGTATT	TAGAGGTCAA	ATGAAAATTC	CATTGTGTGT	3720
	ACAGACCTCA	GAGCTCAGGA	AAGATAAAAAA	AGAATAAATA	AAACTCTAAA	CAGACCTTGA	3780
35	CAAAATTAAT	CCTAGAGACT	GGCACAGACT	TACTTGGTAC	TCCTTCCCCT	TGCCCTATTT	3840
	AGAACTGAGA	ATACTCCCTC	TTGATTGGT	TTTACTCTTT	TTAAGATCCT	TTATGGGCT	3900
	CCTATGCCAT	CACTGTCTTA	AATGATGTGT	TTAAACCTAT	GTTGTTATAA	TAATGATCTA	3960
40	TATGTTAAGT	AAAAGGCTT	GCAGGGTGTG	CAGAAAGAAG	TCTGGTCACA	ACTGGCTACA	4020
	GTGAACAAGC	TGGGTACCCC	AAGGACATCT	TACCAAGTCC	AGCCAGAGAT	CTGATCTACG	4080
45	TACACCTGCG	TCATGCTGAG	ACCCTCAAGC	CTCACTAAAA	GGGTCCCTGC	CTAGTTCTGT	4140
	TTACTAATCT	GCCTTATTCT	GTGTTGTTC	CCATGTTAAA	GATAGAGTAA	ATGCAGTATT	4200
50	CTCCACATAG	AGATATAGAC	TTCTGAAATT	CTAAGATTAG	AATTATTTAC	AAGAAGAAGT	4260
	GGGGAATGAA	GAATAAAAAAA	TTACTGGCCT	CTTGTGAGAA	CATGAACCTT	CACCTCGGAG	4320
	CCCACCCCT	CCCATCTGGA	AAACATACTT	GAGAAAAACA	TTTCTGGAA	CAACCACAGA	4380
55	ATGTTCAAC	AGGCCAGATG	TATTGCCAAA	CACAGGATAT	GACTCTTGG	TTGAGTAAAT	4440
	TTGTGGTTGT	TAAACTCCCTC	CTATTCCCTC	CCCATTCCCC	CTCCCACTTT	GTGGTTTTT	4500
60	CCTTAAAAG	CTTGTGAAAAA	ATTTGAGTCG	TCGTCGAGAC	TCCTCTACCC	TGTGCAAAGG	4560
	TGTATGAGTT	TCGACCCAG	AGCTCTGTGT	GCTTCTGTT	GCTGCTTTAT	TTCGACCCCA	4620
	GAGCTCTGGT	CTGTGTGCTT	TCATGTCGCT	GCTTTATTAA	ATCTTACCTT	CTACATTAA	4680
65	TGTATGGTCT	CAGTGTCTTC	TTGGGTACGC	GGCTGTCCCG	GGACTTGAGT	GTCTGAGTGA	4740

GGGTCTTCCC TCGAGGGTCT TTCATTTGGT ACATGGGCCG GGAATTGAG AATCTTCAT 4800  
 TTGGTGCATT GGCCGGGAAT TCGAAAATCT TTCA 4834

## 5 (2) INFORMATION FOR SEQ ID NO:26:

- (i) SEQUENCE CHARACTERISTICS:
  - (A) LENGTH: 4518 base pairs
  - (B) TYPE: nucleic acid
  - (C) STRANDEDNESS: single
  - (D) TOPOLOGY: linear

15 (ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:26:

20	CACCTGACGC GCCCTGTAGC GGCGCATTA GCGCGCGGG TGTGGTGGTT ACGCGCAGCG	60
	TGACCGCTAC ACTTGCCAGC GCCCTAGCGC CCGCTCCTT CGCTTCTTC CCTTCCTTTC	120
25	TCGCCACGTT CGCCGGCTTT CCCCGTCAAG CTCTAAATCG GGGGCTCCCT TTAGGGTTC	180
	GATTTAGTGC TTTACGGCAC CTCGACCCCA AAAAACATTGA TTAGGGTGAT GGTCACGTA	240
	GTGGGCCATC GCCCTGATAG ACGGTTTTC GCCCTTGAC GTTGGAGTCC ACGTTCTTA	300
30	ATAGTGGACT CTTGTTCCAA ACTGGAACAA CACTCAACCC TATCTGGTC TATTCTTTG	360
	ATTTATAAGG GATTTTGCCT ATTTCGGCCT ATTGGTAAA AAATGAGCTG ATTTAACAAA	420
35	AATTTAACGC GAATTTAAC AAAATATTAA CGCTTACAAT TTACCGTTA AGATACATTG	480
	ATGAGTTGG ACAAAACCACA ACTAGAATGC AGTAAAAAAA ATGCTTATT TGTGAAATT	540
	GTGATGCTAT TGCTTATT GTAACCATT TAAGCTGCAA TAAACAAGTT AACAAACAACA	600
40	ATTGCATTCA TTTTATGTT CAGGTCAGG GGGAGGTGTG GGAGGTTTT TAAAGCAAGT	660
	AAAACCTCTA CAAATGTGGT ATGGCTGATT ATGATCATGA ACAGACTGTG AGGACTGAGG	720
45	GGCCTGAAAT GAGCCTGGG ACTGTGAATC TAAAATACAC AAACAATTAG AATCAGTAGT	780
	TTAACACATT ATACACTTAA AAATTGGATC TCCATTGCC ATTCAAGGCTG CGCAACTGTT	840
	GGGAAGGGCG ATCGGTGCG GCCTCTCGC TATTACGCCA GCTGGCGAAA GGGGGATGTG	900
50	CTGCAAGGCG ATTAAGTTGG GTAACGCCAG GGTTTCCCA GTCACGACGT TGTAAAACGA	960
	CGGCCAGTGA ATTGTAATAC GACTCACTAT AGGGCGAATT GGGTACACTT ACCTGGTACC	1020
55	CCACCCGGGT GGAAAATCGA TGGGCCCGCG GCCGCTCTAG AAGTACTCTC GAGAAGCTTT	1080
	TTGAATTCTT TGGATCCACT AGTGTGACC TGCAAGGCGCG CGAGCTCCAG CTTTGTTCC	1140
	CTTTAGTGAG GGTAAATTTC GAGCTGGCG TAATCAAGGT CATAGCTGTT TCCTGTGTGA	1200
60	AATTGTTATC CGCTCACAAT TCCACACAAT ATACGAGCCG GAAGTATAAA GTGTAAAGCC	1260
	TGGGGTGCCT AATGAGTGAG CTAACTCACA GTAATTGCGG CTAGCGGATC TGACGGTTCA	1320
65	CTAAACCAGC TCTGCTTATA TAGACCTCCC ACCGTACACG CCTACGCCA ATTTGCGTCA	1380
	ATGGGGCGGA GTTGTACGA CATTGGAA AGTCCCGTTG ATTTGGTGC CAAACAAAC	1440

	TCCCCATTGAC	GTCAATGGGG	TGGAGACTTG	GAAATCCCCG	TGAGTCAAAC	CGCTATCCAC	1500
5	GCCCCATTGAT	GTACTGCCAA	AACCGCATCA	CCATGGTAAT	AGCGATGACT	AATACGTAGA	1560
	TGTACTGCCA	AGTAGGAAAG	TCCCATAAGG	TCATGTACTG	GGCATAATGC	CAGGCGGGCC	1620
	ATTTACCGTC	ATTGACGTCA	ATAGGGGGCG	TACTTGGCAT	ATGATACACT	TGATGTACTG	1680
10	CCAAGTGGC	AGTTTACCGT	AAATACTCCA	CCCATTGACG	TCAATGGAAA	GTCCCTATTG	1740
	GCGTTACTAT	GGGAACATAC	GTCATTATTG	ACGTCAATGG	GCGGGGGTGCG	TTGGGCGGTC	1800
15	AGCCAGGCCG	GCCATTAC	GTAAGTTATG	TAACGCGGAA	CTCCATATAT	GGGCTATGAA	1860
	CTAATGACCC	CGTAATTGAT	TACTATTAAT	AACTAATGCA	TGGCGGTAAT	ACGGTTATCC	1920
	ACAGAACATAG	GGGATAACGC	AGGAAAGAAC	ATGTGAGCAA	AAGGCCAGCA	AAAGGCCAGG	1980
20	AACCGTAAAA	AGGCCGCGTT	GCTGGCGTT	TTCCATAGGC	TCCGCCCCCC	TGACGAGCAT	2040
	CACAAAAATC	GACGCTCAAG	TCAGAGGTGG	CGAAACCCGA	CAGGACTATA	AAGATACCAG	2100
	GCGTTTCCCC	CTGGAAGCTC	CCTCGTGC	TCTCCTGTT	CGACCC	TGCTTACCGGA	2160
25	TACCTGTCCG	CCTTCTCCC	TTCGGGAAGC	GTGGCGCTT	CTCATAGCTC	ACGCTGTAGG	2220
	TATCTCAGTT	CGGTGTAGGT	CGTCGCTCC	AAGCTGGGCT	GTGTGCACGA	ACCCCCCGTT	2280
30	CAGCCCGACC	GCTGCGCCTT	ATCCGTAAC	TATCGTCTT	AGTCCAACCC	GGTAAGACAC	2340
	GACTTATCGC	CACTGGCAGC	AGCCACTGGT	AACAGGATTA	GCAGAGCGAG	GTATGTAGGC	2400
	GGTGCTACAG	AGTTCTTGAA	GTGGTGGCCT	AACTACGGCT	ACACTAGAAG	GACAGTATTT	2460
35	GGTATCTGCG	CTCTGCTGAA	GCCAGTTACC	TTCGGAAAAAA	GAGTTGGTAG	CTCTTGATCC	2520
	GGCAAACAAA	CCACCGCTGG	TAGCGGTGGT	TTTTTGTTT	GCAAGCAGCA	GATTACGCGC	2580
40	AGAAAAAAAG	GATCTCAAGA	AGATCCTTTG	ATCTTTCTA	CGGGGTCTGA	CGCTCAGTGG	2640
	AACGAAAAC	CACGTTAAGG	GATTTGGTC	ATGAGATTAT	CAAAAAGGAT	CTTCACCTAG	2700
	ATCCTTTAA	ATTAAAAATG	AAGTTTAAA	TCAATCTAAA	GTATATATGA	GTAACCTGAG	2760
45	GCTATGGCAG	GGCCTGCCGC	CCCGACGTTG	GCTGCGAGCC	CTGGGCCTTC	ACCCGAACCTT	2820
	GGGGGGTGGG	GTGGGGAAAAA	GGAAGAAACG	CGGGCGTATT	GGCCCCAATG	GGGTCTCGGT	2880
50	GGGGTATCGA	CAGAGTGCCA	GCCCTGGGAC	CGAACCCCGC	GTGTATGAAC	AAACGACCCA	2940
	ACACCGTGCG	TTTTATTCTG	TCTTTTATT	GCCGTCTAG	CGCGGGTTCC	TTCCGGTATT	3000
	GTCTCCTTCC	GTGTTTCAGT	TAGCCTCCCC	CTAGGGTGGG	CGAAGAAACTC	CAGCATGAGA	3060
55	TCCCCCGCCT	GGAGGATCAT	CCAGCCGGCG	TCCCGGAAAAA	CGATTCCGAA	GCCCAACCTT	3120
	TCATAGAAGG	CGCGGGTGGG	ATCGAAATCT	CGTGATGGCA	GGTTGGGCGT	CGCTTGGTCG	3180
60	GTCATTTCGA	ACCCCAGAGT	CCCGCTCAGA	AGAACTCGTC	AAGAAGGCGA	TAGAAGGCGA	3240
	TGCGCTGCGA	ATCGGGAGCG	GCGATACCGT	AAAGCACCGAG	GAAGCGGTCA	GCCCATTGCG	3300
	CGCCAAGCTC	TTCAGCAATA	TCACGGGTAG	CCAACGCTAT	GTCCTGATAG	CGGTCCGCCA	3360
65	CACCCAGCCG	GCCACAGTCG	ATGAATCCAG	AAAAGCGGCC	ATTTCCACC	ATGATATTG	3420

	GCAAGCAGGC ATCGCCATGG GTCACCGACGA GATCCTCGCC GTCGGGCATG CTCGCCTTGA	3480
	GCCTGGCGAA CAGTTCGGCT GGCGCGAGCC CCTGATGCTC TTCTGCCAGA TCATCCTGAT	3540
5	CGACAAGACC GGCTTCCATC CGAGTACGTG CTCGCTCGAT GCGATGTTTC GCTTGGTGGT	3600
	CGAATGGGCA GGTAGCCGGA TCAAGCGTAT GCAGCCGCCG CATTGCATCA GCCATGATGG	3660
10	ATACTTCTC GCCAGGAGCA AGGTGAGATG ACAGGAGATC CTGCCCGGC ACTTCGCCCA	3720
	ATAGCAGCCA GTCCCTTCCC GCTTCAGTGA CAACGTCGAG CACAGCTGCG CAAGGAACGC	3780
	CCGTCGTGGC CAGCCACGAT AGCCGCGCTG CCTCGTCTTG CAGTTCATTC AGGGCACCGG	3840
15	ACAGGTCGGT CTTGACAAAA AGAACCGGGC GCCCCTGCGC TGACAGCCGG AACACGGCGG	3900
	CATCAGAGCA GCCGATTGTC TGTTGTGCC AGTCATAGCC GAATAGCCTC TCCACCCAAG	3960
	CGGCCGGAGA ACCTGCGTGC AATCCATCTT GTTCAATCAT GCGAAACGAT CCTCATCCTG	4020
20	TCTCTTGATC GATCTTGCA AAAGCCTAGG CCTCCAAAAA AGCCTCCTCA CTACTTCTGG	4080
	AATAGCTCAG AGGCCGAGGC GGCCTCGGCC TCTGCATAAA TAAAAAAAAT TAGTCAGCCA	4140
	TGGGGCGGAG AATGGGCGGA ACTGGGCGGA GTTAGGGCG GGATGGCGG AGTTAGGGC	4200
25	GGGACTATGG TTGCTGACTA ATTGAGATGC ATGCTTGCA TACTTCTGCC TGCTGGGAG	4260
	CCTGGGGACT TTCCACACCT GGTTGCTGAC TAATTGAGAT GCATGCTTG CATACTTCTG	4320
30	CCTGCTGGGG AGCCTGGGA CTTTCCACAC CCTAACTGAC ACACATTCCA CAGCTGGTTC	4380
	TTTCCGCCCTC AGGACTCTTC CTTTTCAAT ATTATTGAAG CATTATCAG GGTTATTGTC	4440
35	TCATGAGCGG ATACATATT GAATGTATTT AGAAAATAA ACAAAATAGGG GTTCCGCGCA	4500
	CATTTCCCCG AAAAGTGC	4518

## (2) INFORMATION FOR SEQ ID NO:27:

- 40           (i) SEQUENCE CHARACTERISTICS:  
               (A) LENGTH: 25 base pairs  
               (B) TYPE: nucleic acid  
               (C) STRANDEDNESS: single  
               (D) TOPOLOGY: linear
- 45           (ii) MOLECULE TYPE: DNA (genomic)

50           (xi) SEQUENCE DESCRIPTION: SEQ ID NO:27:

CTCCACATAG AGATATAGAC TTCTG

25

## 55           (2) INFORMATION FOR SEQ ID NO:28:

- 60           (i) SEQUENCE CHARACTERISTICS:  
               (A) LENGTH: 45 base pairs  
               (B) TYPE: nucleic acid  
               (C) STRANDEDNESS: single  
               (D) TOPOLOGY: linear
- 65           (ii) MOLECULE TYPE: DNA (genomic)

## (xi) SEQUENCE DESCRIPTION: SEQ ID NO:28:

CGATCTTATT AATTAAC TGG AGTTTGAGC CCRMCCCCTC CCATC

45

## 5 (2) INFORMATION FOR SEQ ID NO:29:

## (i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 5594 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

15 (ii) MOLECULE TYPE: DNA (genomic)

## (xi) SEQUENCE DESCRIPTION: SEQ ID NO:29:

20 TGCATTAGTT ATTAATAGTA ATCAATTACG GGGTCATTAG TTCATAGCCC ATATATGGAG

60

TTCCGCGTTA CATAACTTAC GGTAAATGGC CCGCCTGGCT GACCGCCCAA CGACCCCCGC

120

25 CCATTGACGT CAATAATGAC GTATGTTCCC ATAGTAACGC CAATAGGGAC TTTCCATTGA

180

CGTCAATGGG TGGAGTATT ACAGTAAACT GCCCACTTGG CAGTACATCA AGTGTATCAT

240

ATGCCAAGTA CGCCCCCTAT TGACGTCAAT GACGGTAAAT GGCCCGCCTG GCATTATGCC

300

30 CAGTACATGA CCTTATGGGA CTTTCCTACT TGGCAGTACA TCTACGTATT AGTCATCGCT

360

ATTACCATGG TGATGCGGTT TTGGCAGTAC ATCAATGGC GTGGATAGCG GTTTGACTCA

420

CGGGGATTTC CAAGTCTCCA CCCCATTGAC GTCAATGGGA GTTTGTTTG GCACCAAAAT

480

35 CAACGGGACT TTCCAAAATG TCGTAACAAC TCCGCCCCAT TGACGCAAAT GGGCGGTAGG

540

CGTGTACGGT GGGAGGTCTA TATAAGCAGA GCTGGTTAG TGAACCGTCA GATCCGCGCC

600

40 AGTCCTCCGA TTGACTGAGT CGCCCGGGTA CCCGTGTATC CAATAAACCC TCTTGCAGTT

660

GCATCCGACT TGTGGTCTCG CTGTTCTTG GGAGGGTCTC CTCTGAGTGA TTGACTACCC

720

45 GTCAGGGGG GTCTTCATT TGGGGGCTCG TCCGGGATCG GGAGACCCCT GCCCAGGGAC

780

CACCGACCCA CCACCGGGAG GTAAGCTGGC CAGCAACTTA TCTGTGTCTG TCCGATTGTC

840

TAGTGTCTAT GACTGATT TT ATGCGCCTGC GTCGGTACTA GTTAGCTAAC TAGCTCTGTA

900

50 TCTGGCGGAC CCGTGGTGG ACTGACGAGT TCGGAACACC CGGCCGCAAC CCTGGGAGAC

960

GTCCCAGGAG GAACAGGGGA GGATCAGGGGA CGCCTGGTGG ACCCCTTGAGAGGCCAAGAG

1020

55 ACCATTGGG GTTGCAGAT CGTGGGTTCG AGTCCCACCT CGTCCCCAGT TGCGAGATCG

1080

TGGGTTCGAG TCCCACCTCG TGTGTTGTT CGAGATCGTG GGTTGAGTC CCACCTCGCG

1140

TCTGGTCACG GGATCGTGGG TTCGAGTCCC ACCTCGTGT TTGTTGCGAG ATCGTGGGTT

1200

60 CGAGTCCCAC CTCGCGTCTG GTCACGGGAT CGTGGGTTCG AGTCCCACCT CGTGCAGAGG

1260

GTCTCAATTG GCCGGCCTTA GAGAGGCCAT CTGATTCTTC TGGTTCTCT TTTTGTCTTA

1320

65 GTCTCGTGTGTC CGCTCTTGTGTT GTGACTACTG TTTTCTAAA AATGGGACAA TCTGTGTCCA

1380

CTCCCCTTTC TCTGACTCTG GTTCTGTCGC TTGGTAATT TGTGTTTTA CGTTGTTTT

1440

	TGTGAGTCGT CTATGTTGTC TGTTACTATC TTGTTTTGT TTGTGGTTA CGGTTCTGT	1500
5	GTTATGCCCTT CTAAAATAAG CCTAAAAATC CTGTCAGATC CCTATGCTGA CCACCTCCCT	1560
	TCAGATCAAC AGCTGCCCTT ACGTATCGAT GGATCCCTCG ACTAACTAAT AGCCCATTCT	1620
10	CCAAGGTCGA GCGGGATCAA TTCCGCCCCC CCCCTAACGT TACTGGCCGA AGCCGCTTGG	1680
	AATAAGGCCG GTGTGCGTTT GTCTATATGT TATTTCCAC CATATTGCCG TCTTTGGCA	1740
15	ATGTGAGGGC CCGGAAACCT GGCCCTGTCT TCTTGACGAG CATTCTAGG GGTCTTCCC	1800
	CTCTGCCAA AGGAATGCAA GGTCTGTTGA ATGTCGTGAA GGAAGCAGTT CCTCTGGAAG	1860
	CTTCTTGAAG ACAAAACAACG TCTGTAGCGA CCCTTGCAG GCAGCGGAAC CCCCCACCTG	1920
20	GCGACAGGTG CCTCTGCGGC CAAAGCCAC GTGTATAAGA TACACCTGCA AAGGCGGCAC	1980
	AACCCCAGTG CCACGTTGTG AGTTGGATAG TTGTGGAAAG AGTCAAATGG CTCTCCTCAA	2040
25	GCGTATTCAA CAAGGGGCTG AAGGATGCCA AGAAGGTACC CCATTGTATG GGATCTGATC	2100
	TGGGGCCTCG GTGCACATGC TTTACATGTG TTTAGTCGAG GTTAAAAAAA CGTCTAGGCC	2160
	CCCCGAACCA CGGGGACGTG GTTTCCCTT GAAAAACACG ATAATAATCA TGGCTACAGG	2220
30	CTCCCGGACG TCCCTGCTCC TGGCTTTGG CCTGCTCTGC CTGCCCTGGC TTCAAGAGGG	2280
	CAGTGCCTTC CCAACCATTG CCTTATCCAG GCTTTTGAC AACGCTATGC TCCGCGCCCA	2340
35	TCGTCTGCAC CAGCTGGCCT TTGACACCTA CCAGGAGTTT GAAGAAGCCT ATATCCAAA	2400
	GGAACAGAAAG TATTCAATTCC TGCAGAACCC CCAGACCTCC CTCTGTTCT CAGAGTCTAT	2460
	TCCGACACCC TCCAACAGGG AGGAAACACA ACAGAAATCC AACCTAGAGC TGCTCCGCAT	2520
40	CTCCCTGCTG CTCATCCAGT CGTGGCTGGA GCCCGTGCAG TTCTCAGGA GTGTCTCGC	2580
	CAACAGCCTG GTGTACGGCG CCTCTGACAG CAACGTCTAT GACCTCCTAA AGGACCTAGA	2640
45	GGAAGGCATC CAAACGCTGA TGGGGAGGCT GGAAGATGGC AGCCCCCGGA CTGGGCAGAT	2700
	CTTCAAGCAG ACCTACAGCA AGTCGACAC AAACTCACAC AACGATGACG CACTACTCAA	2760
	GAACTACGGG CTGCTCTACT GCTTCAGGAA GGACATGGAC AAGTCGAGA CATTCCCTGCG	2820
50	CATCGTGCAG TGCCGCTCTG TGGAGGGCAG CTGTGGCTTC TAGCTGCCCG GGTGGCATCC	2880
	TGTGACCCCT CCCCAGTGCC TCTCCTGGCC CTGGAAGTTG CCACTCCAGT GCCCACCAGC	2940
55	CTTGTCTCAA TGTGTGTCAG TTAGGGTGTG GAAAGTCCCC AGGCTCCCCA GCAGGCAGAA	3000
	GTATGCAAAG CATGCATCTC ATTAGTCAG CAACCAGGTG TGGAAAGTCC CCAGGCTCCC	3060
	CAGCAGGGCAG AAGTATGCAA AGCATGCATC TCAATTAGTC AGCAACCATA GTCCCGCCCC	3120
60	TAACTCCGCC CATCCCGCCC CTAACCTCCGC CCAGTTCCGC CCATTCTCCG CCCCATGGCT	3180
	GACTAATTTC TTTTATTAT GCAGAGGCCG AGGCCGCCTC GGCCTCTGAG CTATTCCAGA	3240
65	AGTAGTGAGG AGGCTTTTT GGAGGCCTAG GCTTTGCAA AAAGCTTCAC GCTGCCGCAA	3300
	GCACTCAGGG CGCAAGGGCT GCTAAAGGAA GCGGAACACG TAGAAAGCCA GTCCGCAGAA	3360

	ACGGTGCTGA	CCCCGGATGA	ATGTCAGCTA	CTGGGCTATC	TGGACAAGGG	AAAACGCAAG	3480
5	CGCAAAGAGA	AAGCAGGTAG	CTTGCAGTGG	GCTTACATGG	CGATAGCTAG	ACTGGGCGGT	3540
	TTTATGGACA	GCAAGCGAAC	CGGAATTGCC	AGCTGGGGCG	CCCTCTGGTA	AGGTTGGAA	3600
	GCCCTGCAA	GTAAACTGGA	TGGCTTCTT	GCCGCCAAGG	ATCTGATGGC	GCAGGGGATC	3660
10	AAGATCTGAT	CAAGAGACAG	GATGAGGATC	TTTCGCATG	ATTGAACAAG	ATGGATTGCA	3720
	CGCAGGTTCT	CCGGCCGCTT	GGGTGGAGAG	GCTATTCGGC	TATGACTGGG	CACAACAGAC	3780
	AATCGGCTGC	TCTGATGCCG	CCGTGTTCCG	GCTGTCAGCG	CAGGGGCGCC	CGGTTCTTT	3840
15	TGTCAAGACC	GACCTGTCCG	GTGCCCTGAA	TGAAC TGCA G	GACGAGGCAG	CGCGGCTATC	3900
	GTGGCTGGCC	ACGACGGGCG	TTCCTGCGC	AGCTGTGCTC	GACGTTGTCA	CTGAAGCGGG	3960
20	AAGGGACTGG	CTGCTATTGG	GCGAAGTGCC	GGGGCAGGAT	CTCCTGTCAT	CTCACCTTGC	4020
	TCCTGCCGAG	AAAGTATCCA	TCATGGCTGA	TGCAATGCCG	CGGCTGCATA	CGCTTGATCC	4080
	GGCTACCTGC	CCATTGACCC	ACCAAGCGAA	ACATCGCATC	GAGCGAGCAC	GTACTCGGAT	4140
25	GGAAGCCGGT	CTTGTGATC	AGGATGATCT	GGACGAAGAG	CATCAGGGGC	TCGCGCCAGC	4200
	CGAACTGTTG	GCCAGGCTCA	AGGCGCGCAT	GCCCGACGGC	GAGGATCTCG	TCGTGACCCA	4260
30	TGGCGATGCC	TGCTTGCCGA	ATATCATGGT	GGAAAATGGC	CGCTTTCTG	GATTGATCGA	4320
	CTGTGGCCGG	CTGGGTGTGG	CGGACCGCTA	TCAGGACATA	CGCTGGCTA	CCCGTGATAT	4380
	TGCTGAAGAG	CTTGGCGGCCG	AATGGGCTGA	CCGCTTCCTC	GTGCTTACG	GTATGCCGC	4440
35	TCCCGATTG	CAGCGCATCG	CCTTCTATCG	CCTTCTTGAC	GAGTTCTTCT	GAGCGGGACT	4500
	CTGGGGTTCG	AAATGACCGA	CCAAGCGACG	CCCAACCTCC	AGAAAAAGGG	GGGAATGAAA	4560
40	GACCCCACCT	GTAGGTTGG	CAAGCTAGCT	TAAGTAACGC	CATTTGCAA	GGCATGGAAA	4620
	AATACATAAAC	TGAGAATAGA	GAAGTTCAGA	TCAAGGTCAG	GAACAGATGG	AACAGCTGAA	4680
	TATGGGCCAA	ACAGGATATC	TGTGGTAAGC	AGTCCTGCC	CCGGCTCAGG	GCCAAGAAC	4740
45	GATGGAACAG	CTGAATATGG	GCCAAACAGG	ATATCTGTGG	TAAGCAGTTC	CTGCCCGGC	4800
	TCAGGGCCAA	GAACAGATGG	TCCCCAGATG	CGGTCCAGCC	CTCAGCAGTT	TCTAGAGAAC	4860
50	CATCAGATGT	TTCCAGGGTG	CCCCAAGGAC	CTGAAATGAC	CCTGTGCCTT	ATTTGAACTA	4920
	ACCAATCAGT	TCGCTTCTCG	CTTCTGTTCG	CGCGCTTCTG	CTCCCCGAGC	TCAATAAAAG	4980
	AGCCCACAAAC	CCCTCACTCG	GGGCGCCAGT	AATCTGCTGC	TTGCAAACAA	AAAAACCACC	5040
55	GCTACCAGCG	GTGGTTTGT	TGCCGGATCA	AGAGCTACCA	ACTCTTTTC	CGAAGGTAAC	5100
	TGGCTTCAGC	AGAGCGCAGA	TACCAAATAC	TGTCTTCTA	GTGTAGCCGT	AGTTAGGCCA	5160
60	CCACTTCAAG	AACTCTGTAG	CACCGCCTAC	ATACCTCGCT	CTGCTAATCC	TGTTACCAAGT	5220
	GGCTGCTGCC	AGTGGCGATA	AGTCGTGTCT	TACCGGGTTG	GACTCAAGAC	GATAGTTACC	5280
	GGATAAGGCG	CAGCGGTGG	GCTGAACGGG	GGGTTCGTGC	ACACAGCCCA	GCTTGGAGCG	5340
65	AACGACCTAC	ACCGAACTGA	GATACCTACA	GC GTGAGCAT	TGAGAAAGCG	CCACGCTTCC	5400

CGAAGGGAGA AAGGCGGACA GGTATCCGGT AAGCGGCAGG GTCGGAACAG GAGAGCGCAC	5460
GAGGGAGCTT CCAGGGGGAA ACGCCTGGTA TCTTTATAGT CCTGTCGGGT TTGCCACCT	5520
5 CTGACTTGAG CGTCGATTTC TGTGATGCTC GTCAGGGGGG CGGAGCCTAT GGAAAAACGC	5580
CAGCAACGCC GAGA	5594

## (2) INFORMATION FOR SEQ ID NO:30:

10 (i) SEQUENCE CHARACTERISTICS:  
 (A) LENGTH: 6561 base pairs  
 (B) TYPE: nucleic acid  
 15 (C) STRANDEDNESS: single  
 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

20 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:30:

GATCCCCGGG TCGACCCGGG TCGACCCTGT GGAATGTGTG TCAGTTAGGG TGTGGAAAGT	60
25 CCCCAGGCTC CCCAGCAGGC AGAAGTATGC AAAGCATGCA TCTCAATTAG TCAGCAACCA	120
GGTGTGGAAA GTCCCCAGGC TCCCCAGCAG GCAGAAAGTAT GCAAAGCATG CATCTCAATT	180
30 AGTCAGCAAC CATAGTCCCCG CCCCTAACTC CGCCCATCCC GCCCCTAACT CCGCCCAGTT	240
CCGCCCATTC TCCGCCCAT GGCTGACTAA TTTTTTTAT TTATGCAGAG GCCGAGGCCG	300
35 CCTCGGCCTC TGAGCTATT CAGAAGTAGT GAGGAGGCTT TTTGGAGGC CTAGGCTTT	360
GCAAAAAGCT TCACGCTGCC GCAAGCACTC AGGGCGCAAG GGCTGCTAAA GGAAGCGGAA	420
CACGTAGAAA GCCAGTCCGC AGAACCGGTG CTGACCCGG ATGAATGTCA GCTACTGGC	480
40 TATCTGGACA AGGGAAAACG CAAGCGCAA GAGAAAGCAG GTAGCTTGCA GTGGGCTTAC	540
ATGGCGATAG CTAGACTGGG CGGTTTTATG GACAGCAAGC GAACCGGAAT TGCCAGCTGG	600
45 GGCGCCCTCT GGTAAGGTTG GGAAGCCCTG CAAAGTAAAC TGGATGGCTT TCTTGCCGCC	660
AAGGATCTGA TGGCGCAGGG GATCAAGATC TGATCAAGAG ACAGGATGAG GATCGTTCG	720
CATGATTGAA CAAGATGGAT TGCACGCAGG TTCTCCGGCC GCTTGGGTGG AGAGGCTATT	780
50 CGGCTATGAC TGGGCACAAAC AGACAATCGG CTGCTCTGAT GCCGCCGTGT TCCGGCTGTC	840
AGCGCAGGGG CGCCCGGTTT TTTTGTCAA GACCGACCTG TCCGGTGCC TGAATGAACT	900
55 GCAGGACGAG GCAGCGCGGC TATCGTGGCT GGCCACGACG GGCCTTCCTT GCGCAGCTGT	960
GCTCGACGTT GTCACTGAAG CGGGAAAGGA CTGGCTGCTA TTGGGCGAAG TGCCGGGGCA	1020
GGATCTCCTG TCATCTCACC TTGCTCCTGC CGAGAAAGTA TCCATCATGG CTGATGCAAT	1080
60 GCGGCGGCTG CATAACGCTTG ATCCGGCTAC CTGCCCATTC GACCCACCAAG CGAACACATCG	1140
CATCGAGCGA GCACGTACTC GGATGGAAGC CGGTCTTGTGTC GATCAGGATG ATCTGGACGA	1200
AGAGCATCAG GGGCTCGCGC CAGCCGAACG GTTCGCCAGG CTCAAGGCGC GCATGCCCGA	1260
65 CGCGAGGAT CTCGTCGTGA CCCATGGCGA TGCCTGCTTG CGAAATATCA TGGTGGAAAA	1320

	TGGCCGCTTT TCTGGATTCA TCGACTGTGG CCGGCTGGGT GTGGCggACC GCTATCAGGA	1380
5	CATAGCGTTG GCTACCCGTG ATATTGCTGA AGAGCTTGGC GGCAGAATGGG CTGACCGCTT	1440
	CCTCGTGCTT TACGGTATCG CCGCTCCCGA TTCGCAGCGC ATCGCCTTCT ATCGCCTTCT	1500
	TGACGAGTTC TTCTGAGCGG GACTCTGGGG TTCGAAATGA CCGACCAAGC GACGCCAAC	1560
10	CTGCCATCAC GAGATTCGA TTCCACCGCC GCCTTCTATG AAAGGTTGGG CTTCGGAATC GTTTCCGGG ACGCCGGCTG GATGATCCTC CAGCGCGGGG ATCTCATGCT GGAGTTCTC	1620 1680
	GCCCCACCCCG GAATTCTGAA TCTGCTGCTT GCAAACAAAA AAACCACCGC TACCAGCGGT	1740
15	GGTTTGTGG CGGGATCAAG AGCTACCAAC TCTTTTCCG AAGGTAAC TG GCTTCAGCAG	1800
	AGCGCAGATA CCAAATACTG TCCTTCTAGT GTAGCCGTAG TTAGGCCACC ACTTCAAGAA	1860
20	CTCTGTAGCA CCGCCTACAT ACCTCGCTCT GCTAATCCTG TTACCAAGTGG CTGCTGCCAG	1920
	TGGCGATAAG TCGTGTCTTA CCGGGTTGGA CTCAAGACGA TAGTTACCGG ATAAGGCGCA	1980
	GCGGTCGGGC TGAACGGGGG GTTCGTGCAC ACAGCCCAGC TTGGAGCGAA CGACCTACAC	2040
25	CGAACTGAGA TACCTACAGC GTGAGCATTG AGAAAGCGCC ACGCTTCCCG AAGGGAGAAA	2100
	GGCGGACAGG TATCCGGTAA GCGGCAGGGT CGGAACAGGA GAGCGCACGA GGGAGCTCC	2160
30	AGGGGGAAAC GCCTGGTATC TTTATAGTCC TGTCGGTTT CGCCACCTCT GACTTGAGCG	2220
	TCGATTGGT TGATGCTCGT CAGGGGGCG GAGCCTATGG AAAAACGCCA GCAACGCCGA	2280
	GATGCGCCGC CTCGAGTACA CCTGCGTCAT GCTGAGACCC TCAAGCCTCA CTAAAAGGGT	2340
35	CCCTGCCTAG TTCTGTTTAC TAATCTGCCT TATTCTGTT TTGTTCCCAT GTTAAAGATA	2400
	GAGTAAATGC AGTATTCTCC ACATAGAGAT ATAGACTTCT GAAATTCTAA GATTAGAATT	2460
40	ATTTACAAGA AGAAGTGGGG AATGAAGAAT AAAAATTAC TGGCCTTTG TGAGAACATG	2520
	AACTTTCACC TCGGAGCCCA CCCCCCTCCA TCTGGAAAAC ATACTTGAGA AAAACATTT	2580
	CTGGAACAAAC CACAGAATGT TTCAACAGGC CAGATGTATT GCCAAACACA GGATATGACT	2640
45	CTTTGGTTGA GTAAATTGT GGTTGTTAAA CTTCCCTAT TCCCTCCCCA TTCCCCCTCC	2700
	CAGTTGTGG TTTTTCCCTT TAAAAGCTTG TGAAAAATTG GAGTCGTGCGT CGAGACTCCT	2760
50	CTACCCCTGTG CAAAGGTGTA TGAGTTCGA CCCCAGAGCT CTGTGTGCTT TCTGTTGCTG	2820
	CTTTATTCG ACCCCAGAGC TCTGGCTGT GTGCTTCAT GTCGCTGCTT TATTAAATCT	2880
	TACCTCTAC ATTTATGTA TGGTCTCAGT GTCTTCTGG GTACGCGGCT GTCCCGGGAC	2940
55	TTGAGTGTCT GAGTGAGGGT CTTCCCTCGA GGGTCTTCA TTTGGTACAT GGGCCGGGAA	3000
	TTCGAGAACATC TTTCATTGGA TGCATTGGCC GGGATTGCA AAATCTTCA TTTGGTGCAT	3060
60	TGGCCGGAA ACAGCGCGAC CACCCAGAGG TCCTAGACCC ACTTAGAGGT AAGATTCTT	3120
	GTTCTGTTTT GGTCTGATGT CTGTGTTCTG ATGTCTGTGT TCTGTTCTA AGTCTGGTGC	3180
	GATCGCAGTT TCAGTTTGC GGACGCTCAG TGAGACCGCG CTCCGAGAGG GAGTGCAGGG	3240
65	TGGATAAGGA TAGACGTGTC CAGGTGTCCA CCGTCCGTTC GCCCTGGGAG ACGTCCCAGG	3300

	AGGAACAGGG GAGGATCAGG GACGCCCTGGT GGACCCCTTT GAAGGCCAAG AGACCATTG	3360
	GGGTTGCGAG ATCGTGGGTT CGAGTCCCAC CTCGTGCCA GTTGCAGAT CGTGGGTTCG	3420
5	AGTCCCACCT CGTGTTTGT TGCGAGATCG TGGGTTCGAG TCCCACCTCG CGTCTGGTCA	3480
	CGGGATCGTG GGTTGAGTC CCACCTCGTG TTTGTTGCG AGATCGTGGG TTCGAGTCCC	3540
10	ACCTCGCGTC TGGTCACGGG ATCGTGGGTT CGAGTCCCAC CTCGTGCAGA GGGTCTCAAT	3600
	TGGCCGGCCT TAGAGAGGCC ATCTGATTCT TCTGGTTCT CTTTTGTC TAGTCTCGT	3660
	TCCGCTCTTG TTGTGACTAC TGTTTTCTA AAAATGGGAC AATCTGTGTC CACTCCCCTT	3720
15	TCTCTGACTC TGGTTCTGTC GCTTGGTAAT TTTGTTGTT TACGTTGTT TTTGTGAGTC	3780
	GTCTATGTTG TCTGTTACTA TCTTGTTTT GTTTGTGGTT TACGGTTCT GTGTGTGTCT	3840
	TGTGTGTCTC TTTGTGTTCA GACTTGGACT GATGACTGAC GACTGTTTT AAGTTATGCC	3900
20	TTCTAAAATA AGCCTAAAAAA TCCTGTCAGA TCCCTATGCT GACCACCTCC TTTCAGATCA	3960
	ACAGCTGCC C GCCTCCAC TCCAACCTCA GAGAGCAGCC AGCGGGTCAC AGTGGTCCCG	4020
25	CCCATGAACC TGGAGCCTAG GGAAAAATGA GCTCGGAAAT CCGGAGCAAA TGAGGAGTGG	4080
	TCCCTGAGAA GTCAGTGGCC TAAATGTTGT GGCTGCTGAA GCAAAAGAAG AGGAGGCTGT	4140
	TCGAGTAGCC GCCCAAGAGC GCCGCGGGTT CCCAGGCAGC TTCTCATTCC CCTGTCCCTC	4200
30	CCATCCCGTC TCTTGTAAAC AGAAAAACTG CTTTCACTTT GAGATATGAG TGGCCCGATA	4260
	CAGCCAGCTG TGAGAGCTGT ACTCCCTCC CTGCCAACG TGTTTCTCT TCTCAGGCGA	4320
35	CCCCTCCCTG AGCTGCTGGC AGTGAGTCTG TTCTAAGCTC CAGTGAGGGA GGCATCCGCC	4380
	CACTTGGGC TTCTGTCCAA GTAAAGGAGC ACCTGTGAGT CTAACTGCCA GGCTCTGATG	4440
	GGGGTCTCGT CTCTGTGGGA CTAGAAAGTG TCCCAACAAT CTGACCAAGG TAACAGGAAG	4500
40	TTAAGACAAA GACAGAGACC AAAGTCAGAA TCAGAGCTGT GCTGTGAGAC AAAAGATAA	4560
	AAAAAAATAAA ATGCTGGCCA CAAAGTCAG GAAAAGTAGA AAACCTTAGAT AGTACCTGGC	4620
45	AACAAAAGAA AGCTTTGGC TAAAGATCAA CGTGTATACT GTAAAGAAAA TGAGCACTGG	4680
	GTGAGAGACT GCCCCAACAA AAAGAAGAGG AGCCCCCCTC ATGACCAAAC CCTTCACCTG	4740
	TTCGTGGCTA AAAGTAAAGA GATAACAAAA GGGGTGCTAA CACAGAAGCT GAGTCCTTAA	4800
50	AAGAGTCCGG TGGCCTACCT GTTGAAGCAG CTAAAAAAGA GACTGTGTTT CATACTCCTC	4860
	CACTGACCAG TGCAAAACAA GCTAAAAAGT TCCTGGCAC TGCGGGCTTT TGCAGATTGT	4920
55	GGATTCCAGG TTTTGCTGAG TTAAAGAGAT AAACAGCCCT TCGTATAGAA AAATAAAAAA	4980
	CAACCTTGGA TGTCTTGGA TGCTATTGAG ACTGCCCTAA TGTGTCCTCC AGCTATGGGA	5040
	CTCCTAGATG TGACTGAGAA CAAAGGTATT GCCAAAGAAG TTCTTACTCA GAGATTGGGA	5100
60	CCCTGAAAAA GACCTGTGGC ATACTTGTAA GAAATTAGAC CTGGTGGCTG TAAGATGGCC	5160
	TGCTTGTCTG CACATAGTGG CTTCTGGTCA AGGACGCAGA TAAATTGACT CTGAGACAAA	5220
65	ACTTGGCACA TGTCTTAGAA AGTGTGGTTC AGCCCCCATG ACCGATGGCT GACTAACGCT	5280
	CTTGAAAACA TTATCCAAT GTTCCCTGA CCGATGGACA CATTGTCAGA GCTTTTTTG	5340

	ACTGAACGAG TGACCTTCGC TCCCCCTGCT ATCCTCGATC TCACTACTGC CTGAGACTTC	5400
5	ACCTACTCAT CATTGTGCTG ACATTCTGGC AGAAGAAACT CATACTCGAA ATGATCTGAA	5460
	GGATCAGATC AGCCTTGGCC TGAGAGTTG AGCTGGTACA CGGATGGCAG TAGCCTGGAG	5520
	GTAAAGGGTA AGCGGAAGGC GGGGACAGCA GTGCAGTGGT GGACAGAAAG CAAGTGATCT	5580
	AGGCCAGCAG CCTCCCTAAA GGGACTTCAG CCCACAAAGC CAAACTTGTG GCTTTAACAC	5640
10	AAGCTCTGTA AATGGTAAAAA AAAAAAAAGT CTACACGGAC AGCAGGTATG CTCTTGCCAC	5700
	TGTACAGAGC AATATACAGA CAAAGAGAAC TGTTGACATC TGCAGAGAAA GACCTAAGAT	5760
15	GCTGTGGCTA AAAGAAATCA GATGGCAAAT CTAACCGCCC AGGCATCCTA AAGAGCAATG	5820
	ATCCTGACAG TCTGAAGACT ATCAAGTTAT AGACAAATTA AGACTGGTAA AAAAAACCCCT	5880
	GTATAAAATA GTAAAAACTG AAAAAAGAAA ACTAGTCCTC TCATGAGAAG ACAGACCTGA	5940
20	CATCTACTGA AAAATAGACT TTACTGGAAA AAATATGTGT ATGAATACCT TCTAGTTTT	6000
	GTGAACGTT TCAGAGATGGA TAAAAGCTTT TCCTTGTAAA ACGAGACTGA TCAGATAGTC	6060
25	ATCAAGAAGA TTGTTAAAGA AAATTTCCA AGGTTGGAG TGCCAAAAGC AATAGTGTCA	6120
	GATAATGGTC CTGCCTTTGT TGCCCAGGTA AGTCAGGGTG TGGCCAAGTA TTTAGAGGTC	6180
	AAATGAAAAT TCCATTGTGT GTACAGACCT CAGAGCTCAG GAAAGATAAA AAAGAATAAA	6240
30	TAAAACCTCTA AACAGACCTT GACAAAATTA ATCCTAGAGA CTGGCACAGA CTTACTTGGT	6300
	ACTCCTTCCC CTTGCCCTAT TTAGAACTGA GAATACTCCC TCTTGATTG GTTTACTCT	6360
35	TTTTAAGATC CTTTATGGGG CTCCTATGCC ATCACTGTCT TAAATGATGT GTTTAACCT	6420
	ATGTTGTTAT AATAATGATC TATATGTTAA GTTAAAAGGC TTGCAGGTGG TGCAGAAAGA	6480
	AGTCTGGTCA CAACTGGCTA CAGTGAACAA GCTGGGTACC CCAAGGACAT CTTACCAGTT	6540
40	CCAGCCAGAG ATCTGATCTA C	6561

## (2) INFORMATION FOR SEQ ID NO:31:

- 45           (i) SEQUENCE CHARACTERISTICS:  
               (A) LENGTH: 55 base pairs  
               (B) TYPE: nucleic acid  
               (C) STRANDEDNESS: single  
               (D) TOPOLOGY: linear  
 50           (ii) MOLECULE TYPE: DNA (genomic)

55           (xi) SEQUENCE DESCRIPTION: SEQ ID NO:31:

GACTAACCTT GATTCCACTG GAGCCGTATT ACCGCCATGC ATTAGTTATT AATAG	55
--	----

## 60           (2) INFORMATION FOR SEQ ID NO:32:

- 65           (i) SEQUENCE CHARACTERISTICS:  
               (A) LENGTH: 47 base pairs  
               (B) TYPE: nucleic acid  
               (C) STRANDEDNESS: single  
               (D) TOPOLOGY: linear

5 (ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:32:  
GACTAACCTT GATTCCACTG GAGTAATTGC GGCTAGCGGA TCTGACG

47

10 (2) INFORMATION FOR SEQ ID NO:33:

- (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 66 base pairs  
(B) TYPE: nucleic acid  
15 (C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

20 (ii) MOLECULE TYPE: DNA (genomic)

25 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:33:

GACTAACCTT GATTCCACTG GAGACACTTG ACCTCTACCG CGCCAGTCCT CCGATTGACT  
30 GAGTCG

60

66

30 (2) INFORMATION FOR SEQ ID NO:34:

- (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 48 base pairs  
(B) TYPE: nucleic acid  
35 (C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

40 (ii) MOLECULE TYPE: DNA (genomic)

45 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:34:

GACTAACCTT GATTCCACTG GAGGGATCCG CGCCCATGAT TATTATCG

48

45 (2) INFORMATION FOR SEQ ID NO:35:

- (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 55 base pairs  
(B) TYPE: nucleic acid  
50 (C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

55 (ii) MOLECULE TYPE: DNA (genomic)

60 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:35:

GACTAACCTT GATTCCAGCA ATGTCATGGC TACAGGGTCC CGGACGTCCC TGCTC

55

65 (2) INFORMATION FOR SEQ ID NO:36:

- (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 48 base pairs

- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

5           (ii) MOLECULE TYPE: DNA (genomic)

10           (xi) SEQUENCE DESCRIPTION: SEQ ID NO:36:

GACTAACCTT GATTCCAGCA ATGTTAGGAC AAGGCTGGTG GGCACCTGG

48

(2) INFORMATION FOR SEQ ID NO:37:

- 15           (i) SEQUENCE CHARACTERISTICS:
- (A) LENGTH: 49 base pairs
  - (B) TYPE: nucleic acid
  - (C) STRANDEDNESS: single
  - (D) TOPOLOGY: linear

20           (ii) MOLECULE TYPE: DNA (genomic)

25           (xi) SEQUENCE DESCRIPTION: SEQ ID NO:37:

GACTAACCTT GATTCCACTG GAGGGTCGAC CCTGTGGAAT GTGTGTCAG

49

30           (2) INFORMATION FOR SEQ ID NO:38:

- 35           (i) SEQUENCE CHARACTERISTICS:
- (A) LENGTH: 48 base pairs
  - (B) TYPE: nucleic acid
  - (C) STRANDEDNESS: single
  - (D) TOPOLOGY: linear

40           (ii) MOLECULE TYPE: DNA (genomic)

45           (xi) SEQUENCE DESCRIPTION: SEQ ID NO:38:

48           GACTAACCTT GATTCCACTG GAGAATCTCG TGATGGCAGG TTGGGCAGT

48

(2) INFORMATION FOR SEQ ID NO:39:

- 50           (i) SEQUENCE CHARACTERISTICS:
- (A) LENGTH: 54 base pairs
  - (B) TYPE: nucleic acid
  - (C) STRANDEDNESS: single
  - (D) TOPOLOGY: linear

55           (ii) MOLECULE TYPE: DNA (genomic)

60           (xi) SEQUENCE DESCRIPTION: SEQ ID NO:39:

GACTAACCTT GATTCCACTG AAGAGATTTT ATTTAGTCTC CAGAAAAAGG GGGG

54

(2) INFORMATION FOR SEQ ID NO:40:

65           (i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 50 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

5

(ii) MOLECULE TYPE: DNA (genomic)

10

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:40:

GACTAACCTT GATTCCACTG AAGCCCCAA ATGAAAGACC CCCGCTGACG

50

15 (2) INFORMATION FOR SEQ ID NO:41:

- (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 49 base pairs  
(B) TYPE: nucleic acid  
20 (C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

25

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:41:

30 GACTAACCTT GATTCCACTG GAGCCGGGAC GGAATTCTGTA ATCTGCTGC

49

(2) INFORMATION FOR SEQ ID NO:42:

- (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 47 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
35 (D) TOPOLOGY: linear

40

(ii) MOLECULE TYPE: DNA (genomic)

45

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:42:

GACTAACCTT GATTCCACTG GAGTTCTCGA GGCGGCGCAT CTCGGCG

47

50

(2) INFORMATION FOR SEQ ID NO:43:

- (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 20 base pairs  
(B) TYPE: nucleic acid  
55 (C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

60

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:43:

CGCTCTAGAA CTAGTGGATC

20

65

(2) INFORMATION FOR SEQ ID NO:44:

5                   (i) SEQUENCE CHARACTERISTICS:  
                      (A) LENGTH: 21 base pairs  
                      (B) TYPE: nucleic acid  
                      (C) STRANDEDNESS: single  
                      (D) TOPOLOGY: linear

10                  (ii) MOLECULE TYPE: DNA (genomic)

15                  (xi) SEQUENCE DESCRIPTION: SEQ ID NO:44:

15                  GTAATACGAC TCACTATAGG G

21

15                  (2) INFORMATION FOR SEQ ID NO:45:

20                   (i) SEQUENCE CHARACTERISTICS:  
                      (A) LENGTH: 43 base pairs  
                      (B) TYPE: nucleic acid  
                      (C) STRANDEDNESS: single  
                      (D) TOPOLOGY: linear

25                  (ii) MOLECULE TYPE: DNA (genomic)

30                  (xi) SEQUENCE DESCRIPTION: SEQ ID NO:45:

30                  CGATCCACTG GAGCTCGGAG CCCACCCCT CCCATCTAGA GGT

43

35                  (2) INFORMATION FOR SEQ ID NO:46:

35                   (i) SEQUENCE CHARACTERISTICS:  
                      (A) LENGTH: 43 base pairs  
                      (B) TYPE: nucleic acid  
                      (C) STRANDEDNESS: single  
                      (D) TOPOLOGY: linear

40                  (ii) MOLECULE TYPE: DNA (genomic)

45

45                  (xi) SEQUENCE DESCRIPTION: SEQ ID NO:46:

50                  CGTCCTCCTG GAGAGCACAG GGTAGAGGAG TCTCGACGGT CAG

43

50                  (2) INFORMATION FOR SEQ ID NO:47:

55                   (i) SEQUENCE CHARACTERISTICS:  
                      (A) LENGTH: 43 base pairs  
                      (B) TYPE: nucleic acid  
                      (C) STRANDEDNESS: single  
                      (D) TOPOLOGY: linear

60                  (ii) MOLECULE TYPE: DNA (genomic)

65

65                  (xi) SEQUENCE DESCRIPTION: SEQ ID NO:47:

65                  CGCAACCCTG GAGACCTCTA GATGGGAGGG GGTGGGCTCC GAG

43

(2) INFORMATION FOR SEQ ID NO:48:

5 (i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 43 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

10 (ii) MOLECULE TYPE: DNA (genomic)

15 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:48:

GCAGGACCTG GAGCTGACCG TCGAGACTCC TCTACCCTGT GCT

43

20 (2) INFORMATION FOR SEQ ID NO:49:

25 (i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 20 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

30 (ii) MOLECULE TYPE: DNA (genomic)

35 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:49:

CGCTCTAGAA CTAGTGGATC

20

40 (2) INFORMATION FOR SEQ ID NO:50:

45 (i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 21 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

50 (ii) MOLECULE TYPE: DNA (genomic)

55 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:50:

59 GTAATACGAC TCACTATAGG G

21

60 (2) INFORMATION FOR SEQ ID NO:51:

65 (i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 19 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

65 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:51:

TACGTATCGA TGGATCCGA

19

(2) INFORMATION FOR SEQ ID NO:52:

- 5 (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 19 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

10 (ii) MOLECULE TYPE: DNA (genomic)

15 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:52:

GGATCCATCG ATACGTAAG

19

20 (2) INFORMATION FOR SEQ ID NO:53:

- 25 (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 38 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

30 (ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:53:

35 GGCGGCTAAC TAATAGCCC TTCTCCAAGG TACGTAGC

38

(2) INFORMATION FOR SEQ ID NO:54:

- 40 (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 38 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

45 (ii) MOLECULE TYPE: DNA (genomic)

50 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:54:

TACGTACCTT GGAGAATGGG CTATTAGTTA GCGGCCGC

38

55 (2) INFORMATION FOR SEQ ID NO:55:

- 55 (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 55 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

60 (ii) MOLECULE TYPE: DNA (genomic)

65

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:55:  
GACTAACCTT GATTCCACTG GAGTTTCTC TATTCTTCAT TCCCCACTTC TTCTT

55

## (2) INFORMATION FOR SEQ ID NO:56:

5

- (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 60 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

10

(ii) MOLECULE TYPE: DNA (genomic)

15

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:56:

GACTAACCTT GATTCCACTG GAGAATCTGG ACCAATTCTA TATAAGCCTG TGAAAAATT

60

20

## (2) INFORMATION FOR SEQ ID NO:57:

25

- (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 46 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

30

(ii) MOLECULE TYPE: DNA (genomic)

35

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:57:

GACTAACCTT GATTCCACTG GAGAAGAAGA AGTGGGAAAT GAAGAA

46

40

## (2) INFORMATION FOR SEQ ID NO:58:

45

- (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 51 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

50

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:58:

GACTAACCTT GATTCCACTG GAGATCTCTA GATGGGAGGG GGTCTGGGCT C

51

55

## (2) INFORMATION FOR SEQ ID NO:59:

60

- (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 47 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

65

(ii) MOLECULE TYPE: DNA (genomic)

5 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:59:

GACTAACCTT GATTCCACTG GAGCTCGGAG CCCACCCCT CCCATCT

47

10 (2) INFORMATION FOR SEQ ID NO:60:

15 (i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 47 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

15 (ii) MOLECULE TYPE: DNA (genomic)

20 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:60:

GACTAACCTT GATTCCACTG GAGGGAGGCC CTTATCTCAA AAATGTT

47

25 (2) INFORMATION FOR SEQ ID NO:61:

30 (i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 51 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

30 (ii) MOLECULE TYPE: DNA (genomic)

35

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:61:

GACTAACCTT GATTCCACTG GAGTCTAAGA ACATTTTG AATAAGGGCC T

51

40 (2) INFORMATION FOR SEQ ID NO:62:

45 (i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 44 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

50 (ii) MOLECULE TYPE: DNA (genomic)

50

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:62:

55 GACTAACCTT GATTCCACTG GAGTCACAGG CTTATATAGT GAAA

44

60 (2) INFORMATION FOR SEQ ID NO:63:

60 (i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 46 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

65 (ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:63:

5 GACTAACCTT GATTCCCTGG AGACTGCACT GCTGTCCCCG CCTTCG

46

(2) INFORMATION FOR SEQ ID NO:64:

10 (i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 53 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

15 (ii) MOLECULE TYPE: DNA (genomic)

20 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:64:

GAGTAACCTT GATTCCCTGG AGATTTCTCA GACCCGGGTC GACCCTGTGG AAT

53

25 (2) INFORMATION FOR SEQ ID NO:65:

25 (i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 44 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

30 (ii) MOLECULE TYPE: DNA (genomic)

35

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:65:

40 GACTAACCTT GATTCCCTGG AGCTCGAGGC GGCGCATCTC GGCG

44

40 (2) INFORMATION FOR SEQ ID NO:66:

45 (i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 47 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

50 (ii) MOLECULE TYPE: DNA (genomic)

55 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:66:

55 GACTAACCTT GATTCCCTGA AGACCTGCGT CATGCTGAGA CCCTCAA

47

(2) INFORMATION FOR SEQ ID NO:67:

60 (i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 50 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

65 (ii) MOLECULE TYPE: DNA (genomic)

5 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:67:

GACTAACCTT GATTCCCTGA AGCGGCCAAT GCACCAAATG AAAGATTTTC

50

(2) INFORMATION FOR SEQ ID NO:68:

10 (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 50 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

15 (ii) MOLECULE TYPE: DNA (genomic)

20 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:68:

CGCATTTT AATTAAGTGG AGARAATTG TYACAGGCTT ATATAGKAAA

50

We claim:

1. A method for assembling a gene or gene vector comprising the steps of:

- 5           a) designing at least 6 primers to produce at least three fragments in at least three separate polymerase chain reactions wherein each primer comprises at least one predetermined restriction endonuclease recognition site that recognizes a restriction endonuclease that cleaves at a distance from the recognition site, a sequence complementary to a template sequence for amplification, and bases positioned at the restriction endonuclease  
10 cleavage site that are selected to be complementary to only one other overhanging created from enzymatic cleavage of the fragments;
- 15           b) combining the primers with template nucleic acid and performing a gene amplification reaction to produce multiple copies of an amplified template fragment incorporating the restriction endonuclease recognition site;
- 15           c) digesting the amplified template fragments with one or more restriction endonucleases that recognize the restriction endonuclease recognition site of the primers to create overhanging termini wherein each overhanging termini is complementary to only one other overhanging termini on another fragment; and
- 20           d) combining the amplified and digested template fragments in a ligation reaction to produce a directionally ordered gene, nucleic acid fragment or gene vector.

2. The method of claim 1 wherein the restriction endonuclease is at least one class IIS restriction endonuclease.

- 25           3. The method of claim 2 wherein the class IIS restriction endonuclease is selected from the group consisting of: *AlwI*, *Alw26I*, *BbsI*, *BbvI*, *BbvII*, *BpmI*, *BsmAI*, *BsmI*, *BsmBI*, *BspMI*, *BsrI*, *BsrDI*, *Eco57I*, *EarI*, *FokI*, *GsuI*, *HgaI*, *HphI*, *MboII*, *MnII*, *PleI*, *SapI*, *SfaNI*, *TaqII*, *Tth111II*.
- 30           4. The method of claim 1 wherein class II restriction endonuclease recognition sites, linkers, or adapters are not used to create the gene or gene vector.

5. The method of claim 1 wherein the product of the ligation reaction is introduced into prokaryotic or eukaryotic cells.

5 6. The method of claim 1 wherein at least one target nucleic acid sequence is chosen from the group consisting of : transcriptional regulatory sequences; genetic vectors; introns and/or exons; viral encapsidation sequences; integration signals intended for introducing nucleic acid molecules into other nucleic acid molecules; retrotransposon(s); VL30 elements; or multiple allelic forms of a sequence.

10

7. The method of claim 1 wherein the method is used to generate combinatorial libraries of a target sequence.

8. The method of claim 7 wherein the target sequence is part or all of a gene.

15

9. The method of claim 8 wherein the gene encodes a protein.

10. The method of claim 8 wherein the primers amplify allelic variants of part or all of a gene.

20

11. The method of claim 1 wherein the product of the ligation reaction is passed between eukaryotic cells using a virus particle, by cell fusion, or by transfection.

25

12. The method of claim 1 wherein the product of the ligation reaction is not introduced into prokaryotic cells.

13. The method of claim 1 further combining at least one screening or selection step to select the products of the ligation reaction.

30 14. The method of claim 1 wherein the product of the ligation reaction is mutated during passage in cells in order to generate genetic diversity.

15. The method of claim 14 wherein the product of the ligation reaction is mutated by homologous recombination during passage in cells.

16. The method of claim 1, wherein the method is used to isolate and identify regulatory  
5 sequences from a cell.

17. The method of claim 11, wherein cells containing the product of the ligation reaction are selected for enhanced biological activity.

10 18. The method of claim 17, wherein the cells containing the product of the ligation reaction are selected for tissue-specific, hormone-specific or developmental-specific gene expression.

15 19. The method of claim 1 wherein the product of the ligation reaction is a circularized gene vector.

20. A nucleic acid primer having a 5' and a 3' end to amplify a nucleic acid fragment for the ligation of at least two fragments comprising:

20 a restriction endonuclease recognition site that recognizes a restriction endonuclease, wherein the restriction endonuclease cleaves at a distance from the recognition site and creates overhanging termini;

a sequence complementary to a template sequence to be amplified to produce the nucleic acid fragment;

25 at least two nucleic acid bases positioned at the restriction endonuclease cleavage site and that form an overhanging terminus after cleavage by the restriction endonuclease, wherein the at least two nucleic acid bases are selected to be complementary to only one other overhanging terminus on another fragment of the ligation; and

an affinity handle on the 5' end of the primer.

30 21. The primer of claim 20 further comprising an anchor to provide stability to the restriction enzyme at the restriction enzyme recognition site.

22. A method for isolating and identifying promoters comprising the steps of:

- a) obtaining a vector comprising at least a portion of a promoter region from a retrovirus transposon LTR and having two non-complementary overhanging termini;
- b) designing at least two PCR primers to amplify at least one region of a retro-transposon LTR from template nucleic acid to produce at least one nucleic acid fragment wherein each primer comprises at least one predetermined restriction endonuclease recognition site that recognizes a restriction endonuclease that cleaves at a distance from the recognition site, a sequence complementary to a template sequence from a retrovirus transposon, and bases positioned at the restriction endonuclease cleavage site that are selected to be complementary to only one other overhanging terminus of the vector wherein the restriction endonuclease cleavage site is created from enzymatic cleavage of the fragments;
- c) combining the primers with template nucleic acid and performing a gene amplification reaction to produce multiple copies of an amplified template fragment incorporating the restriction endonuclease recognition site;
- d) digesting the amplified template fragments with one or more restriction endonuclease that recognize the restriction endonuclease recognition site of the primer to create overhanging termini; and
- e) combining the amplified and digested template fragment in a ligation reaction with the vector to produce a gene vector with an intact LTR sequence.

20

23. The method of claim 22 wherein the template nucleic acid is DNA or RNA.

24. The method of claim 22 further comprising the step of sequencing the insert to identify the promoter sequence.

25

25. Promoter sequences of SEQ ID NOS:2-13 identified using the methods of claim 22.

26. The vector of SEQ ID NO:1.

30

1/9

Fig 1A

A.

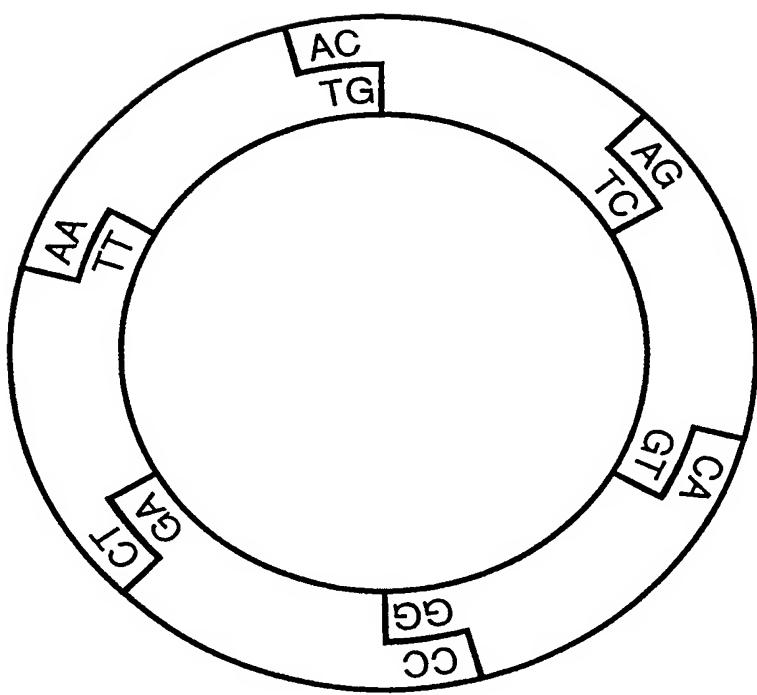
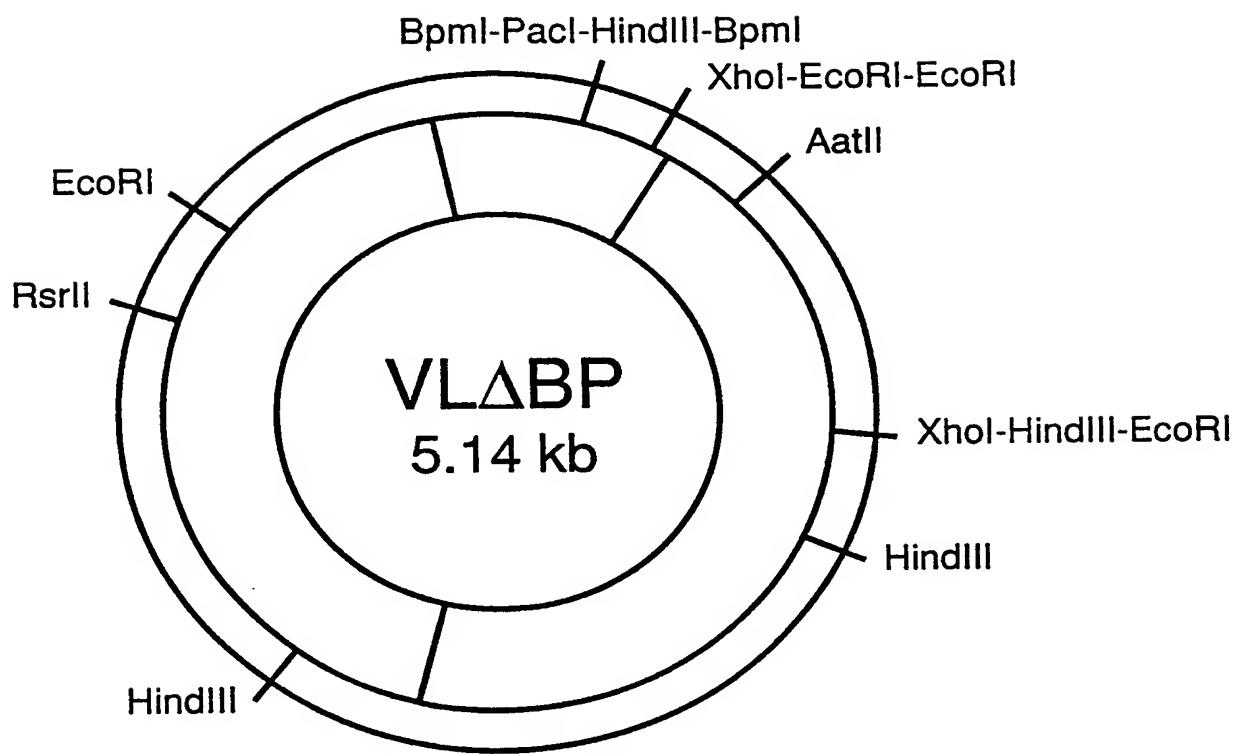


Fig 1B

B.



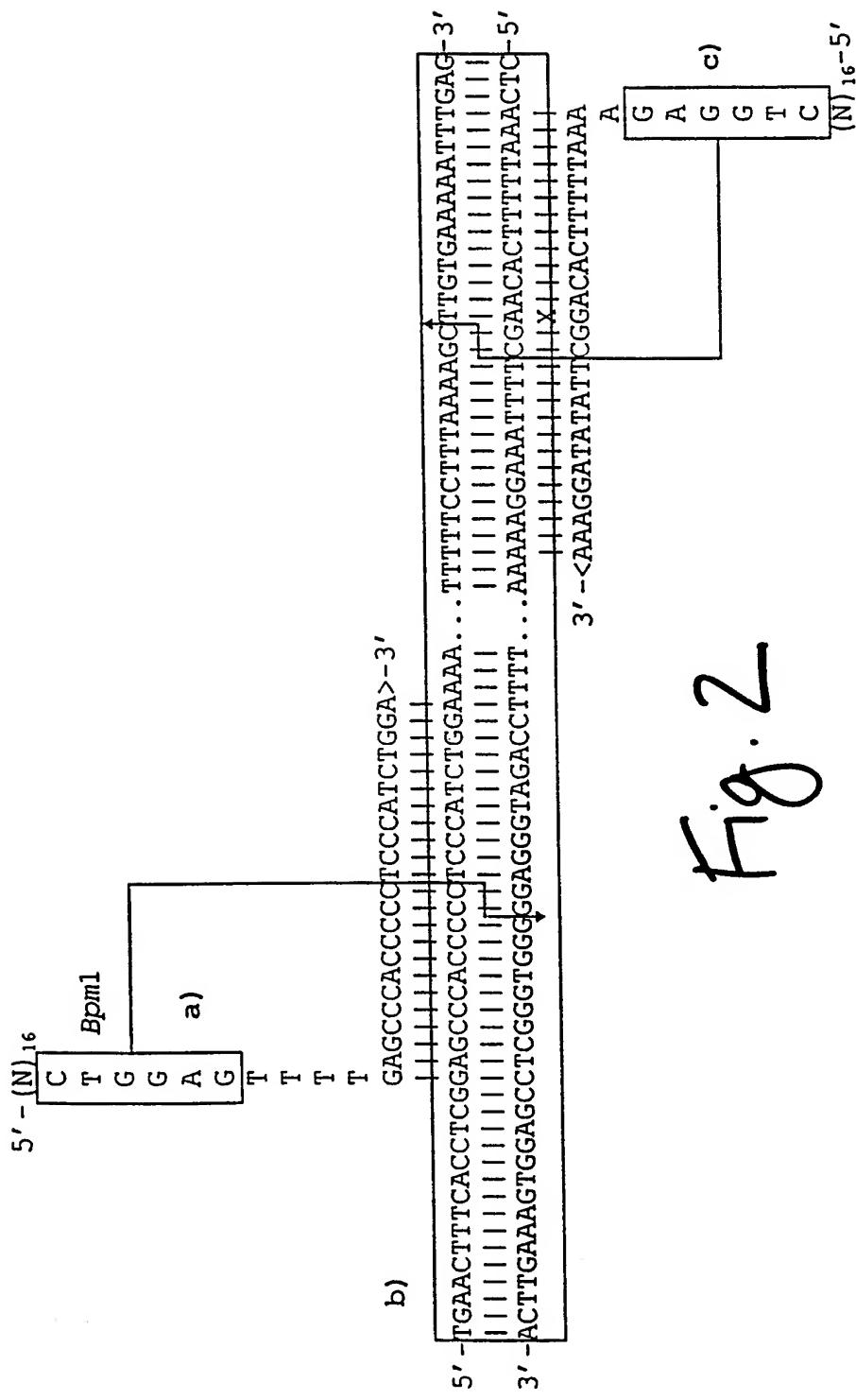


Fig. 2

Fig. 3

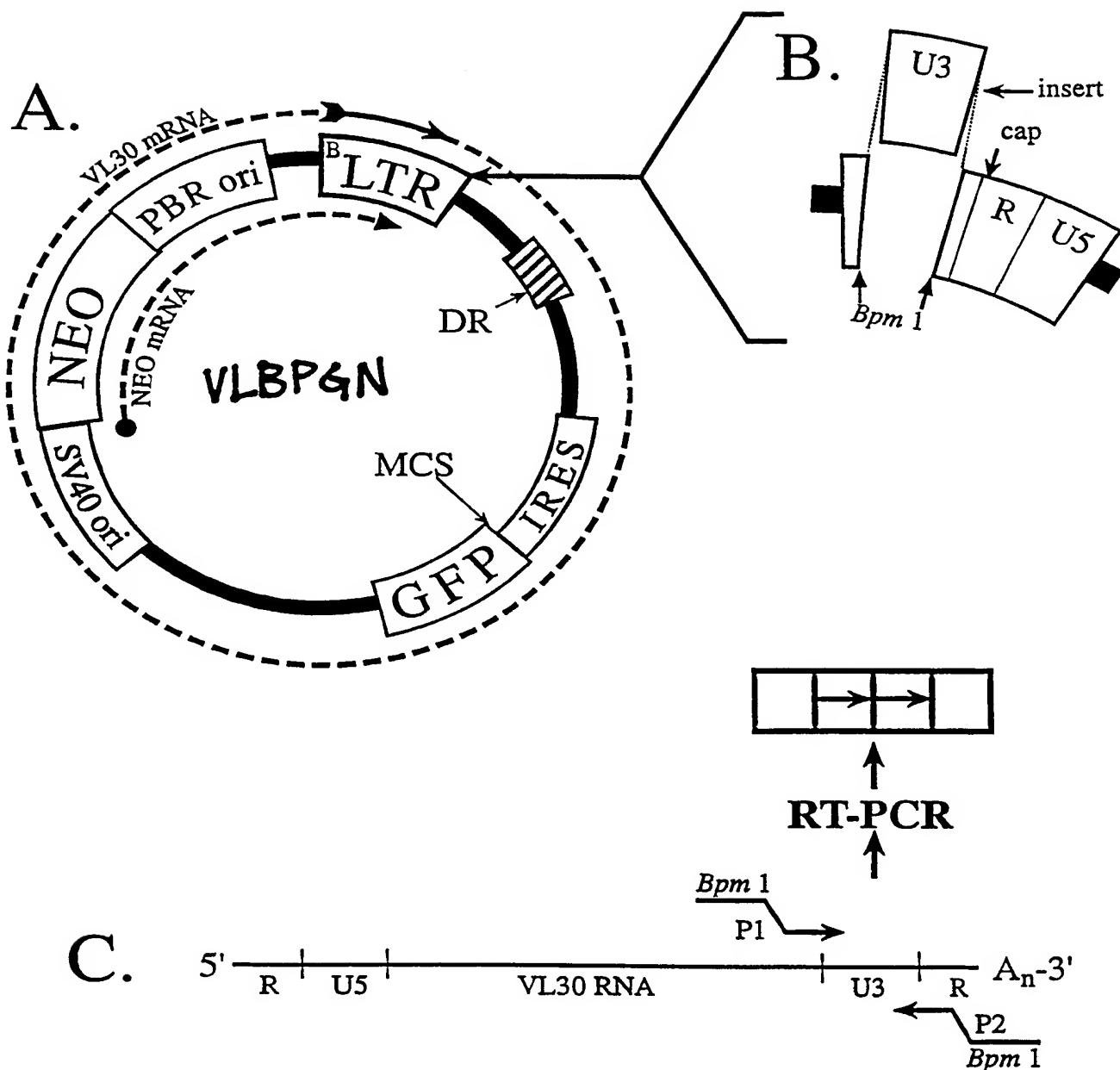


Fig. 4 A.

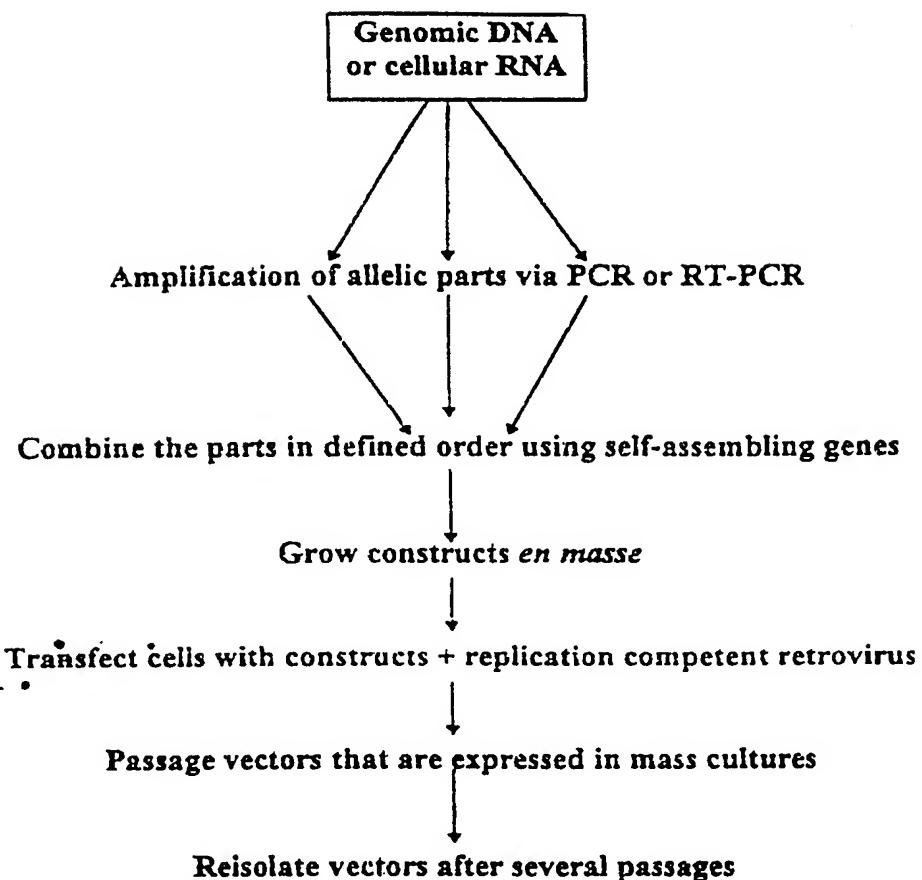


Fig 4 B.

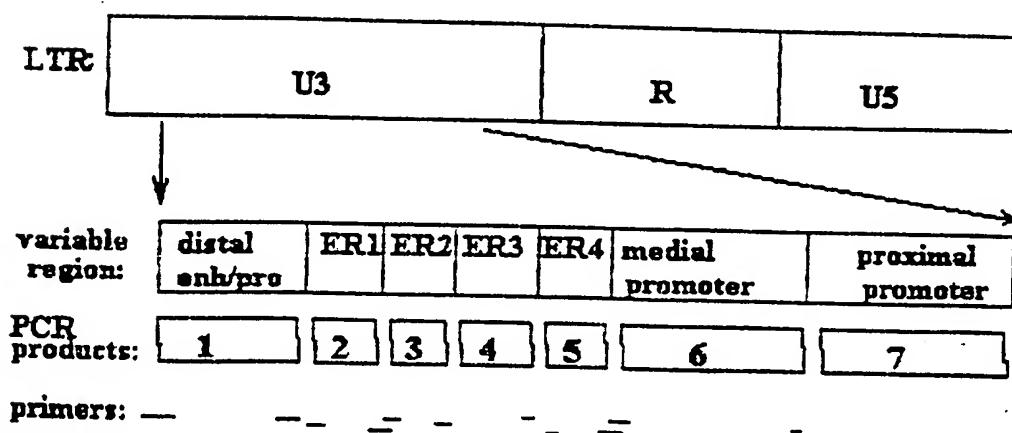


Fig. 5

I.D.2	CCTCCCCATCT AGAGGTGTT CTCGGAACAC TCCCTAAACTT TTCAACCCAA AACTCCTCAC	CCPAAGTC GAAAAGTC CCTCAAGAC	90
I.D.3	-----	-----	-----
I.D.2	ATTTTCTAGA TAAAGGCCTC CTAGGAAAC CTCAAAATGA CATTGCCAA TGATAGACA	TGACTCCCTTA GTTACGAGG TMCCTTGATA	180
I.D.3	-----	-----	-----
I.D.2	GGACATRACT CCTTAGTTAC GTAGGTTCT TGATAGACA TGATAGACA TAAAGGCTC	TGATAGACA TGATAGACA TGATAGACA TAAAGGCTC	270
I.D.3	ACCAGGTACT CCTTACGTTAC GTAGGTTCT TGATAGACA TGATAGACA TGATAGACA	TGATAGACA TGATAGACA TGATAGACA TAAAGGCTC	321
I.D.2	ACTTGACTT TCCCTGCCA GTTCCTCCCC TTAGAGTTT ACTATATAAG C	AGAACTCCCT AGGAGATPA	
I.D.3	ACTTGACTT TCCCTGCCA GTTCCTCCCC TTAGAGTTT ACTATATAAG C	AGAACTCCCT AGGAGATPA	

Fig. 6

Fig. 7

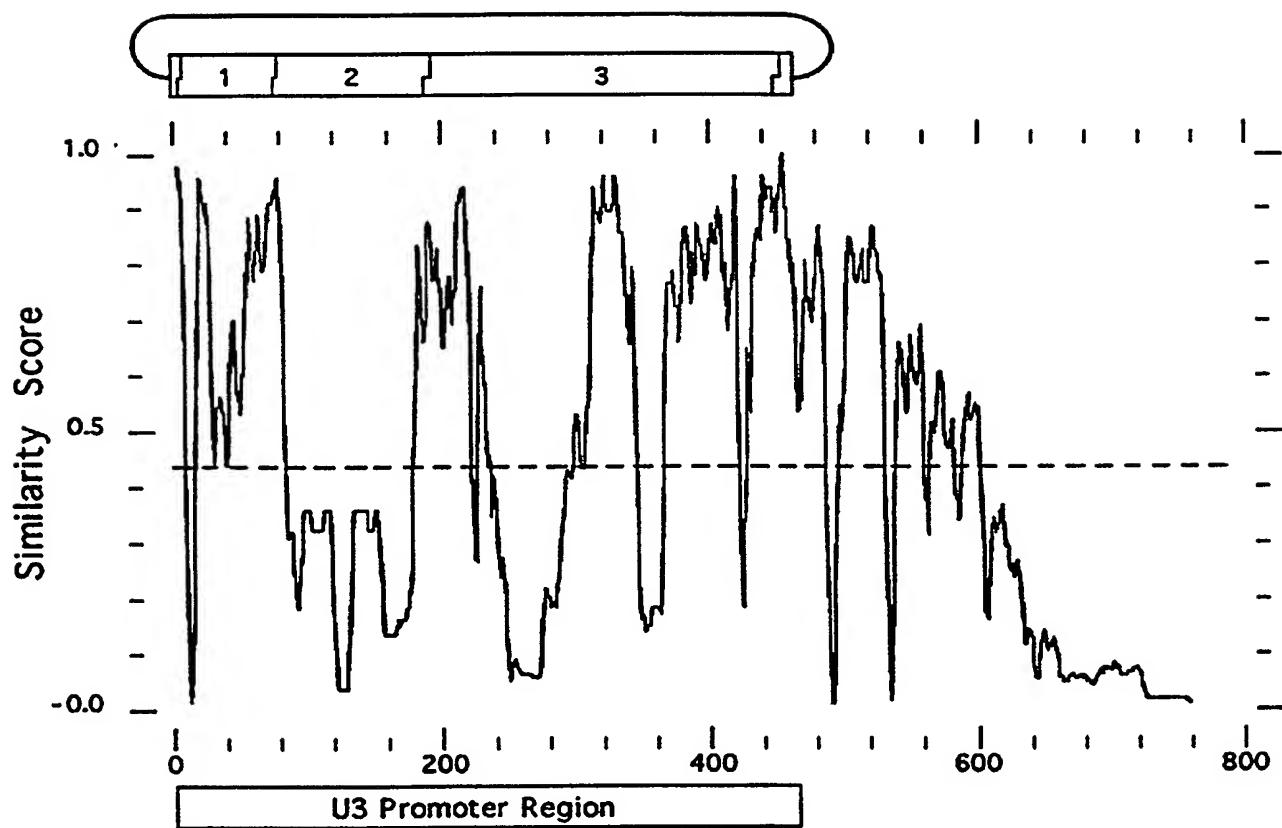
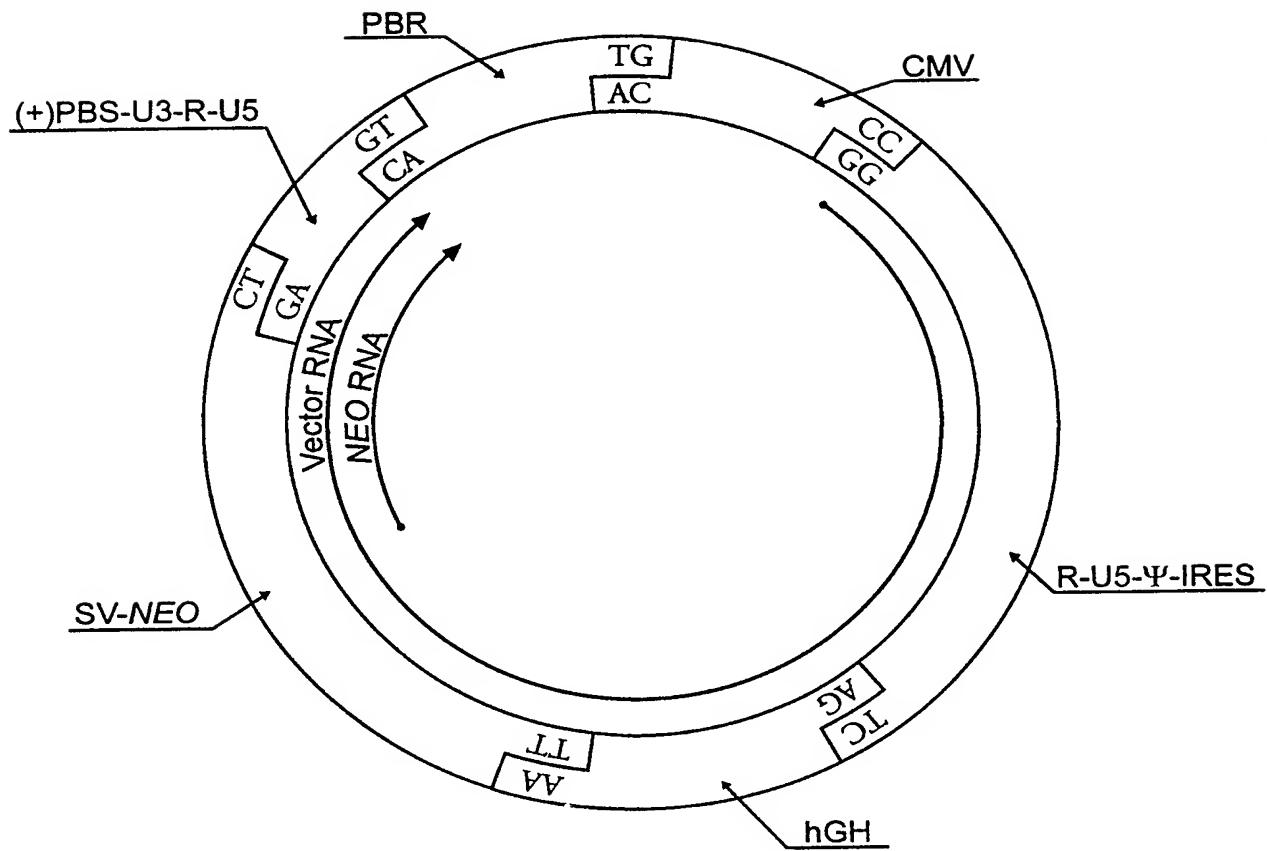


Fig. 8



# INTERNATIONAL SEARCH REPORT

I.	International Application No PCT/US 98/03918
----	---

**A. CLASSIFICATION OF SUBJECT MATTER**  
 IPC 6 C12N15/86 C12N15/10

According to International Patent Classification (IPC) or to both national classification and IPC

**B. FIELDS SEARCHED**

Minimum documentation searched (classification system followed by classification symbols)

IPC 6 C12N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

Category	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	PADGETT K A ET AL: "Creating seamless junctions independent of restriction sites in PCR cloning" GENE, vol. 168, no. 1, 2 February 1996, page 31-35 XP004042930 see the whole document ---	1,2, 4-14, 19-21
Y	---	3
Y	TOMIC, M. ET AL.: "A rapid and simple method for introducing specific mutations into any position of DNA leaving all other positions unaltered" NUCLEIC ACIDS RESEARCH, vol. 18, no. 6, 1990, OXFORD GB, page 1656 XP002069445 cited in the application see the whole document ---	3
	-/--	

Further documents are listed in the continuation of box C.

Patent family members are listed in annex.

**Special categories of cited documents :**

- "A" document defining the general state of the art which is not considered to be of particular relevance
- "E" earlier document but published on or after the international filing date
- "L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
- "O" document referring to an oral disclosure, use, exhibition or other means
- "P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search

Date of mailing of the international search report

26 June 1998

09/07/1998

Name and mailing address of the ISA

European Patent Office, P.B. 5818 Patentlaan 2  
NL - 2280 HV Rijswijk  
Tel. (+31-70) 340-2040, Tx. 31 651 epo nl.  
Fax: (+31-70) 340-3016

Authorized officer

Chambonnet, F

# INTERNATIONAL SEARCH REPORT

In national Application No

PCT/US 98/03918

## C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	LEBEDENKO, E.N. ET AL.: "Method of artificial DNA splicing by directed ligation" NUCLEIC ACIDS RESEARCH, vol. 19, no. 24, 1991, OXFORD GB, pages 6757-6761, XP002069446 cited in the application see the whole document ---	1
A	CHAKRABORTY, A.K. ET AL.: "Synthetic retrotransposon vectors for gene therapy" FASEB JOURNAL., vol. 7, no. 10, July 1993, FOR EXPERIMENTAL BIOLOGY, BETHESDA, MD US, pages 971-977, XP002029486 see the whole document ---	1
P,X	WO 97 28282 A (STRATAGENE INC) 7 August 1997	1.2, 4-14, 19-21
P,Y	see the whole document ---	3
P,X	HODGSON, C.P. ET AL.: "Self-assembling genes (SAGE) : construction of complex vectors and combinatorial libraries without sub-cloning" CANCER GENE THERAPY, vol. 4, no. 6 conf. suppl., November 1997, page s27 XP002069448 see the whole document ---	1
P,X	ZINK, A. M. ET AL.: "Transcriptional targeting with rescued LTRs : a hepatocyte promoter" CANCER GENE THERAPY, vol. 4, no. 6 conf. suppl., November 1997, page s28 XP002069449 see the whole document -----	22

# INTERNATIONAL SEARCH REPORT

Information on patent family members

International Application No

PCT/US 98/03918

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 9728282	A 07-08-1997	NONE	